region of 16S rDNA, and differ in other regions (variable regions) of the 16S rRNA. These differences can be exploited to allow identification of the different subtype strains. The full sequence of 16S ribosomal RNA or DNA read from the chip is compared against a database of the sequence of thousands of known pathogens to type unambiguously most nonviral pathogens infecting AIDS patients.

In a further embodiment, the invention provides chips which also contain probes for detection of bacterial genes conferring antibiotic resistance. An antibiotic resistance gene can be detected by hybridization to a single probe employed in a reverse dot blot format. Alternatively, a group of probes can be designed according to the same principles discussed above to read all or part the DNA sequence encoding an antibiotic resistance gene. Analogous probes groups are designed for reading other antibiotic resistance gene sequences. Antibiotic resistance frequently resides in one of the following genes in microorganisms coparasitizing AIDS patients: rpoB (encoding RNA polymerase), katG (encoding catalase peroxidase, and DNA gyrase A and B genes.

The inclusion of probes for combinations of tests on a single chip simulates the clinical diagnosis tree that a physician would follow based on the presentation of a given syndrome which could be caused by any number of possible pathogens. Such chips allow identification of the presence and titer of HIV in a patient, identification of the HIV strain type and drug resistance, identification of opportunistic pathogens, and identification of the drug resistance of such pathogens. Thus, the physician is simultaneously apprised of the full spectrum of pathogens infecting the patient and the most effective treatments therefor.

Exemplary HIV Chips

(a) HV 273

The HV 273 chip contains an array of oligonucleotide probes for analysis of an 857 base HIV amplicon between nucleotides 2090 and 2946 (HIVBRU strain numbering). The chip contains four groups of probes: 11 mers, 13 mers, 15 mers and 17 mers. From top to bottom, the HV 273 chip is occupied by rows of 11 mers, followed by rows of 13 mers, followed by rows of 15 mers followed by rows of 17 mers. The interrogation position is nucleotide 6, 7, 8 and 9 respectively in the different sized chips. This arrangement of the different sized probes is referred to as being "in series." Within each size group, there are four probe sets laid down in an A-lane, a C-lane a G-lane and a T-lane respectively. Each lane contains an overlapping series of probes with one probe for each nucleotide in the 2090-2946 HIV reverse transcriptase reference sequence. (*i.e.*, 857 probes per lane). The lanes also include a few column positions which are empty or occupied by control probes. These positions serve to orient the chip, determine background fluorescence and punctuate different subsequences within the target. The chip has an area of 1.28 x 1.28 cm, within which the probes form a 130 X 135 matrix (17,550 cells total). The area occupied by each probe (*i.e.*, a probe cell) is about 98 X 95 microns.

The chip was tested for its capacity to sequence a reverse transcriptase fragment from the HIV strain SF2. An 831 bp RNA fragment (designated pPol19) spanning most of the HIV reverse transcriptase coding sequence was amplified by PCR, using primers tagged with T3 and T7 promoter sequences. The primers, designated RT#1-T3 and 89-391 T7 are shown in Table 4; *see also* Gingeras et al., *J. Inf. Dis.* 164, 1066-1074 (1991) (incorporated by reference in its entirety for all purposes). RNA was labelled by incorporation of fluorescent nucleotides. The RNA was fragmented by heating and hybridized to the chip for 40 min at 30 degrees. Hybridization signals were quantified by fluorescence imaging.

Taking the best data from the four probes sets at each position in the target sequence, 715 out of 821 bases were

read correctly (87%). (Comparisons are based on the sequence of pPol19 determined by the conventional dideoxy method to be identical to SF2). In general, the longer sized probes yielded more sequence than the shorter probes. Of the 21 positions at which the SF2 and BRU strains diverged within the target, 19 were read correctly.

Many of the short ambiguous regions in the target arise in segments of the target flanking the points at which the SF2 and BRU sequences diverge. These ambiguities arise because in these regions the comparison of hybridization signals is not drawn between perfectly matched and single base mismatch probes but between a single-mismatched probe and three probes having two mismatches. These ambiguities in reading an SF2 sequence would not detract from the chip's ability to read a BRU sequence either alone or in a mixture with an SF2 target sequence.

In a variation of the above procedure, the chip was treated with RNase after hybridization of the pPol19 target to the probes. Addition of RNase digests mismatched target and thereby increases the signal to noise ratio. RNase treatment increased the number of correctly read bases to 743/821 or 90% (combining the data from the four groups of probes).

In a further variation, the RNA target was replaced with a DNA target containing the same segment of the HIV genome. The DNA probe was prepared by linear amplification using Taq polymerase, RT#1-T3 primer, and fluorescein d-UTP label. The DNA probe was fragmented with uracil DNA glycosylase and heat treatment. The hybridization pattern across the array and percentage of readable sequence were similar to those obtained using an RNA target. However, there were a few regions of sequence that could be read from the RNA target that could not be read from the DNA target and vice versa.

(b) HV 407 Chip

The 407 chip was designed according to the same principles as the HV 273 chip, but differs in several respects. First, the oligonucleotide probes on this chip are designed to exhibit perfect sequence identity (with the

exception of the interrogation position on each probe) to the HIV strain SF2 (rather than the BRU strain as was the case for the HV 273 chip). Second, the 407 chip contains 13 mers, 15 mers, 17 mers and 19 mers (with interrogation positions at nucleotide 7, 8, 9 and 10 respectively), rather than the 11 mers, 13 mers, 15 mers and 17 mers on the HV 273 chip. Third, the different sized groups of oligomers are arranged in parallel in place of the in-series arrangement on the HV 273 chip. In the parallel arrangement, the chip contains from top to bottom a row of 13 mers, a row of 15 mers, a row of 17 mers, a row of 19 mers, followed by a further row of 13 mers, a row of 15 mers, a row of 17 mers, a row of 19 mers, followed by a row of 13 mers, and so forth. Each row contains 4 lanes of probes, an A lane, a C lane, a G lane and a T lane, as described above. The probes in each lane tile across the reference sequence. The layout of probes on the HV 407 chip is shown in Fig. 10.

The 407 chip was separately tested for its ability to sequence two targets, pPol19 RNA and 4MUT18 RNA. pPol19 contains an 831 bp fragment from the SF2 reverse transcriptase gene which exhibits perfect complementarity to the probes on the 407 chip (except of course for the interrogation positions in three of the probes in each column). 4MUT18 differs from the reference sequence at thirty-one positions within the target, including five positions in codons 67, 70, 215 and 219 associated with acquisition of drug resistance. Target RNA was prepared, labelled and fragmented as described above and hybridized to the HV 407 chip. The hybridization pattern for the pPol19 target is shown in Fig. 11.

The sequences read off the chip for the pPol19 and 4MUT18 targets are both shown in Fig. 12 (although the two sequences were determined in different experiments). The sequence labelled wildtype in the Figure is the reference sequence. The four lanes of sequence immediately below the reference sequence are the respective sequences read from the four-sized groups of probes for the pPol19 target (from top-to-bottom, 13 mers, 15 mers, 17 mers and 19 mers). The next four lanes of sequence are the sequences read from the four-sized groups of

probes for the 4MUT18 target (from top-to-bottom in the same order). The regions of sequences shown in normal type are those that could be read unambiguously from the chip. Regions where sequence could not be accurately read are shown highlighted. Some regions of sequence that could not be read from one sized set of probes could be read from another.

Taking the best result from the four sized groups of probes at each column position, about 97% of bases in the pPol19 sequence and about 90% of bases in the 4MUT18 sequence were read accurately. Of the 31 nucleotide differences between 4MUT18 and the reference sequence, twenty-seven were read correctly including three of the nucleotide changes associated with acquisition of drug resistance. Of the ambiguous regions in the 4MUT18 sequence determination, most occurred in the 4MUT18 segments flanking points of divergence between the 4MUT18 and reference sequences. Notably, most of the common mutations in HIV reverse transcriptase associated with drug resistance (*see* Table 3) occur at sequence positions that can be read from the chip. Thus, most of the commonly occurring mutations can be detected by a chip containing an array of probes based on a single reference sequence.

Comparison of the sequence read of the probes of different sizes is useful in determining the optimum size probe to use for different regions of the target. The strategy of customizing probe length within a single group of probe sets minimizes the total number of probes required to read a particular target sequence. This leaves ample capacity for the chip to include probes to other reference sequences (*e.g.*, 16S RNA for pathogenic microorganisms) as discussed below.

The HV 407 chip has also been tested for its capacity to detect mixtures of different HIV strains. The mixture comprises varying proportions of two target sequences; one a segment of a reverse transcriptase gene from a wildtype SF2 strain, the other a corresponding segment from an SF2 strain bearing a codon 67 mutation. *See* Fig. 13. The Figure also represents the probes on the chip having an interrogation position for reading the nucleotide in which the mutation

occurs. A single probe in the Figure represents four probes on the chip with the symbol (o) indicating the interrogation position, which differs in each of the four probes. Figure 14 shows the fluorescence intensity for the four 13 mers and the four 15 mers having an interrogation position for reading the nucleotide in the target sequence in which the mutation occurs. As the percentage of mutant target is increase, the fluorescence intensity of the probe exhibiting perfect complementarity to the wildtype target decreases, and the intensity of the probe exhibiting perfect complementarity to the mutant sequence increases. The intensities of the other two probes do not change appreciably. It is concluded that the chip can be used to analyze simultaneously a mixture of strains, and that a strain comprising as little as ten percent of a mixture can be easily detected.

c. Protease Chip

A protease chip was constructed using the basic tiling strategy. The chip comprises four probes tiling across a 382 nucleotide span including 297 nucleotides from the protease coding sequence. The reference sequence was a consensus Clay-B HIV protease sequence. Different probes lengths were employed for tiling different regions of the reference sequence. Probe lengths were 11, 14, 17 and 20 nucleotides with interrogation positions at or adjacent to the center of each probe. Lengths were optimized from prior hybridization data employing a chip having multiple tilings, each with a different probe length.

The chip was hybridized to four different single-stranded DNA protease target sequences (HXB2, SF2, NY5, pPol4mut18). Both sense and antisense strands were sequenced. Data from the chip was compared with that from an ABI sequencer. The overall accuracy from sequencing the four targets is illustrated in the Table 5 below.

Table 5

	ABI		Protease Chip	
	Sense	Antisense	Sense	Antisense
No call	0	4	9	4
Ambiguous	6	14	17	8
Wrong call	2	3	3	1
TOTAL	8	21	29	13

ABI (sense) - 99.5%
Chip (sense) - 98.1%

ABI (antisense) - 98.6%
Chip (antisense) - 99.1%

Combining the data from sense and antisense strands, both the chip and the ABI sequencer provided 100% accurate data for all of the sequence from all four clones.

In a further test, the chip was hybridized to protease target sequences from viral isolates obtained from four patients before and after ddI treatment. The sequence read from the chip is shown in Fig. 15. Several mutations (indicated by arrows) have arisen in the samples obtained posttreatment. Particularly noteworthy was the chip's capacity to read a g/a mutation at nucleotide 207, notwithstanding the presence of two additional mutations (gt) at adjacent positions.

B. Cystic Fibrosis Chips

A number of years ago, cystic fibrosis, the most common severe autosomal recessive disorder in humans, was shown to be associated with mutations in a gene thereafter named the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene. The CFTR gene is about 250 kb in size and has 27 exons. Wildtype genomic sequence is available for all exonic regions and exons/intron boundaries (Zielenski et al., *Genomics* 10, 214-228 (1991). The full-length wildtype cDNA sequence has also been described (*see* Riordan et al., *Science* 245, 1059-1065 (1989). Over 400 mutations have been mapped (*see* Tsui et al, *Hu. Mutat.* 1, 197-203 (1992). Many of the more common mutations are shown in Table 6. The most common cystic

fibrosis mutation is a three-base deletion resulting in the omission of amino acid #508 from the CFTR protein. The frequency of mutations varies widely in populations of different geographic or ethnic origin (*see* column 4 of Table 6). About 90% of all mutations having phenotypic effects occur in coding regions.

Detection of CFTR mutations is useful in a number of respects. For example, screening of populations can identify asymptomatic heterozygous individuals. Such individuals are at risk of giving rise to affected offspring suffering from CF if they reproduce with other such individuals. *In utero* screening of fetuses is also useful in identifying fetuses bearing 2 CFTR mutations. Identification of such mutations offers the possibility of abortion, or gene therapy. For couples known to be at risk of giving rise to affected progeny, diagnosis can be combined with *in vitro* reproduction procedures to identify an embryo having at least one wildtype CF allele before implantation. Screening children shortly after birth is also of value in identifying those having 2 copies of the defective gene. Early detection allows administration of appropriate treatment (*e.g.*, Pulmozyme Antibiotics, Pertussive Therapy) thereby improving the quality of life and perhaps prolonging the life expectancy of an individual.

The source of target DNA for detecting of CFTR mutations is usually genomic. In adults, samples can conveniently be obtained from blood or mouthwash epithelial cells. In fetuses, samples can be obtained by several conventional techniques such as amniocentesis, chorionic villus sampling or fetal blood sampling. At birth, blood from the amniotic chord is a useful tissue source.

The target DNA is usually amplified by PCR. Some appropriate pairs of primers for amplifying segments of DNA including the sites of known mutations are listed in Tables 5 and 6.

Table 7

OLIGO NUMBER	SEQUENCE
787	TCTCCTTGGATATACTTGTGTGAATCAA
788	TCACCAGATTTCGTAGTCTTTTCATA
851	GTCTTGTGTTGAAATTCTCAGGGTAT
769	CTTGTACCAGCTCACTACCTAAT
887	ACCTGAGAAGATAGTAAGCTAGATGAA
888	AACTCCGCCTTTCCAGTTGTAT
934	TTAGTTTCTAGGGGTGGAAGATACA
935	TTAATGACACTGAAGATCACTGTTCTAT
789	CCATTCCAAGATCCCTGATATTTGAA
790	GCACATTTTTGCAAAGTTCATTAGA
891	TCATGGGCCATGTGCTTTTCAA
892	ACCTTCCAGCACTACAAACTAGAA
760	CAAGTGAATCCTGAGCGTGATTT
850	GGTAGTGTGAAGGGTTCATATGCATA
762	GATTACATTAGAAGGAAGATGTGCCTTT
763	ACATGAATGACATTTACAGCAAATGCTT
931	GTGACCATATTGTAATGCATGTAGTGA
932	ATGGTGAACATATTTCTCAAGAGGTAA
955	TGT CTC TGT AAA CTG ATG GCT AAC A
884	TCGTATAGAGTTGATTGGATTGAGAA
885	CCATTAACTTAATGTGGTCTCATCACAA
886	CTACCATAATGCTTGGGAGAAATGAA
782	TCAAAGAATGGCACCAGTGTGAAA
901	TGCTTAGCTAAAGTTAATGAGTTCAT

OLIGO NUMBER	SEQUENCE
784	AATTGTGAAATTGTCTGCCATTCTTAA
785	GATTCACTTACTGAACACAGTCTAACAA
791	AGGCTTCTCAGTGATCTGTTG
792	GAATCATTCAGTGGGTATAAGCA
1013	GCCATGGTACCTATATGTCACAGAA
1012	TGCAGAGTAATATGAATTTCTTGAGTACA
766	GGGACTCCAAATATTGCTGTAGTAT
1065	GTACCTGTTGCTCCAGGTATGTT

Other primers can be readily devised from the known genomic and cDNA sequences of CFTR. The selection of primers, of course, depends on the areas of the target sequence that are to be screened. The choice of primers also depends on the strand to be amplified. For some regions of the CFTR gene, it makes little difference to the hybridization signal whether the coding or noncoding strand is used. In other regions, one strand may give better discrimination in hybridization signals between matched and mismatched probes than the other. The upper limit in the length of a segment that can be amplified from one pair of PCR primers is about 50 kb. Thus, for analysis of mutants through all or much of the CFTR gene, it is often desirable to amplify several segments from several paired primers. The different segments may be amplified sequentially or simultaneously by multiplex PCR. Frequently, fifteen or more segments of the CFTR gene are simultaneously amplified by PCR. The primers and amplifications conditions are preferably selected to generate DNA targets. An asymmetric labelling strategy incorporating fluorescently labelled dNTPs for random labelling and dUTP for target fragmentation to an average length of less than 60 bases is preferred. The use of dUTP and fragmentation with

uracil N-glycosylase has the added advantage of eliminating carry over between samples.

Mutations in the CFTR gene can be detected by any of the tiling strategies noted above. The block tiling strategy is one particularly useful approach. In this strategy, a group (or block) of probes is used to analyze a short segment of contiguous nucleotides (*e.g.*, 3, 5, 7 or 9) from a CFTR gene centered around the site of a mutation. The probes in a group are sometimes referred to as constituting a block because all probes in the group are usually identical except at their interrogation positions. As noted above, the probes may also differ in the presence of leading or trailing sequences flanking regions of complementary. However, for ease of illustration, it will be assumed that such sequences are not present. As an example, to analyze a segment of five contiguous nucleotides from the CFTR gene, including the site of a mutation (such as one of the mutations in Table 6), a block of probes usually contains at least one wildtype probe and five sets of mutant probes, each having three probes. The wildtype probe has five interrogation positions corresponding to the five nucleotides being analyzed from the reference sequence. However, the identity of the interrogation positions is only apparent when the structure of the wildtype probe is compared with that of the probes in the five mutant probe sets. The first mutant probe set comprises three probes, each being identical to the wildtype probe, except in the first interrogation position, which differs in each of the three mutant probes and the wildtype probe. The second through fifth mutant probe sets are similarly composed except that the differences from the wildtype probe occur in the second through fifth interrogation position respectively. Note that in practice, each set of mutant probes is sometimes laid down on the chip juxtaposed with an associated wildtype probe. In this situation, a block would comprise five wildtype probes, each effectively providing the same information. However, visual inspection and confidence analysis of the chip is facilitated by the largely redundant information provided by five wildtype probes.

After hybridization to labelled target, the relative hybridization signals are read from the probes. Comparison of the intensities of the three probes in the first mutant probe set with that of the wildtype probe indicates the identity of the nucleotide in the target sequence corresponding to the first interrogation position. Comparison of the intensities of the three probes in the second mutant probe set with that of the wildtype probe indicates the identity of the nucleotide in the target sequence corresponding to the second interrogation position, and so forth. Collectively, the relative hybridization intensities indicate the identity of each of the five contiguous nucleotides in the reference sequence.

In a preferred embodiment, a first group (or block) of probes is tiled based on a wildtype reference sequence and a second group is tiled based a mutant version of the wildtype reference sequence. The mutation can be a point mutation, insertion or deletion or any combination of these. The combination of first and second groups of probes facilitates analysis when multiple target sequences are simultaneously applied to the chip, as is the case when a patient being diagnosed is heterozygous for the CFTR allele.

The above strategy is illustrated in Fig. 16, which shows two groups of probes tiled for a wildtype reference sequence and a point mutation thereof. The five mutant probe sets for the wildtype reference sequence are designated wt1-5, and the five mutant probe sets for the mutant reference sequence are designated m1-5. The letter N indicates the interrogation position, which shifts by one position in successive probe sets from the same group. The figure illustrates the hybridization pattern obtained when the chip is hybridized with a homozygous wildtype target sequence comprising nucleotides n-2 to n+2, where n is the site of a mutation. For the group of probes tiled based on the reference sequence, four probes are compared at each interrogation position. At each position, one of the four probes exhibits a perfect match with the target, and the other three exhibit a single-base mismatch. For the group of probes tiled based on the mutant

reference sequence, again four probes are compared at each interrogation position. At position, n, one probe exhibits a perfect match, and three probes exhibit a single base mismatch. Hybridization to a homozygous mutant yields an analogous pattern, except that the respective hybridization patterns of probes tiled on the wildtype and mutant reference sequences are reversed.

The hybridization pattern is very different when the chip is hybridized with a sample from a patient who is heterozygous for the mutant allele (*see* Fig. 17). For the group of probes tiled based on the wildtype sequence, at all positions but n, one probe exhibits a perfect match at each interrogation position, and the other three probes exhibit a one base mismatch. At position n, two probes exhibit a perfect match (one for each allele), and the other probes exhibit single-base mismatches. For the group of probes tiled on the mutant sequence, the same result is obtained. Thus, the heterozygote point mutant is easily distinguished from both the homozygous wildtype and mutant forms by the identity of hybridization patterns from the two groups of probes.

Typically, a chip comprises several paired groups of probes, each pair for detecting a particular mutation. For example, some chips contain 5, 10, 20, 40 or 100 paired groups of probes for detecting the corresponding numbers of mutations. Some chips are customized to include paired groups of probes for detecting all mutations common in particular populations (*see* Table 6). Chips usually also contain control probes for verifying that correct amplification has occurred and that the target is properly labelled.

The goal of the tiling strategy described above is to focus on short regions of the CTFR region flanking the sites of known mutation. Other tiling strategies analyze much larger regions of the CFTR gene, and are appropriate for locating and identifying hitherto uncharacterized mutations. For example, the entire genomic CFTR gene (250 kb) can be tiled by the basic tiling strategy from an array of about one million probes. Synthesis and scanning of such an array of probes is entirely feasible. Other tiling strategies, such as

the block tiling, multiplex tiling or pooling can cover the entire gene with fewer probes. Some tiling strategies analyze some or all of components of the CFTR gene, such as the cDNA coding sequence or individual exons. Analysis of exons 10 and 11 is particularly informative because these are location of many common mutations including the ΔF508 mutation.

Exemplary CFTR chips

One illustrative chip bears an array of 1296 probes covering the full length of exon 10 of the CFTR gene arranged in a 36 x 36 array of 356 μm elements. The probes in the array can have any length, preferably in the range of from 10 to 18 residues and can be used to detect and sequence any single-base substitution and any deletion within the 192-base exon, including the three-base deletion known as ΔF508. As described in detail below, hybridization of nanomolar concentrations of wild-type and ΔF508 oligonucleotide target nucleic acids labeled with fluorescein to these arrays produces highly specific signals (detected with confocal scanning fluorescence microscopy) that permit discrimination between mutant and wild-type target sequences in both homozygous and heterozygous cases.

Sets of probes of a selected length in the range of from 10 to 18 bases and complementary to subsequences of the known wild-type CFTR sequence are synthesized starting at a position a few bases into the intron on the 5'-side of exon 10 and ending a few bases into the intron on the 3'-side. There is a probe for each possible subsequence of the given segment of the gene, and the probes are organized into a "lane" in such a way that traversing the lane from the upper left-hand corner of the chip to the lower righthand corner corresponded to traversing the gene segment base-by-base from the 5'-end. The lane containing that set of probes is, as noted above, called the "wild-type lane."

Relative to the wild-type lane, a "substitution" lane, called the "A-lane", was synthesized on the chip. The A-lane probes were identical in sequence to an adjacent (immediately below the corresponding) wild-type probe but contained, regardless of the sequence of the wild-type probe, a dA

residue at position 7 (counting from the 3'-end). In similar fashion, substitution lanes with replacement bases dC, dG, and dT were placed onto the chip in a "C-lane," a "G-lane," and a "T-lane," respectively. A sixth lane on the chip consisted of probes identical to those in the wild-type lane but for the deletion of the base in position 7 and restoration of the original probe length by addition to the 5'-end the base complementary to the gene at that position.

The four substitution lanes enable one to deduce the sequence of a target exon 10 nucleic acid from the relative intensities with which the target hybridizes to the probes in the various lanes. Various versions of such exon 10 DNA chips were made as described above with probes 15 bases long, as well as chips with probes 10, 14, and 18 bases long. For the results described below, the probes were 15 bases long, and the position of substitution was 7 from the 3'-end.

The sequences of several important probes are shown below. In each case, the letter "X" stands for the interrogation position in a given column set, so each of the sequences actually represents four probes, with A, C, G, and T, respectively, taking the place of the "X." Sets of shorter probes derived from the sets shown below by removing up to five bases from the 5'-end of each probe and sets of longer probes made from this set by adding up to three bases from the exon 10 sequence to the 5'-end of each probe, are also useful and provided by the invention.

3'-TTTATAXTAGAAACC
3'- TTATAGXAGAAACCA
3'- TATAGTXGAAACCAC
3'- ATAGTAXAAACCACA
3'- TAGTAGXAACCACAA
3'- AGTAGAXACCACAAA
3'- GTAGAAXCCACAAAG
3'- TAGAAAXCACAAAGG
3'- AGAAACXACAAAGGA

To demonstrate the ability of the chip to distinguish the ΔF508 mutation from the wild-type, two synthetic target

nucleic acids were made. The first, a 39-mer complementary to a subsequence of exon 10 of the CFTR gene having the three bases involved in the ΔF508 mutation near its center, is called the "wild-type" or wt508 target, corresponds to positions 111-149 of the exon, and has the sequence shown below:

5'-CATTAAAGAAAATATCATCTTTGGTGTTTCCTATGATGA.

The second, a 36-mer probe derived from the wild-type target by removing those same three bases, is called the "mutant" target or mu508 target and has the sequence shown below, first with dashes to indicate the deleted bases, and then without dashes but with one base underlined (to indicate the base detected by the T-lane probe, as discussed below):

5'-CATTAAAGAAAATATCAT---TGGTGTTTCCTATGATGA;

5'-CATTAAAGAAAATATCATTGGTGTTTCCTATGATGA.

Both targets were labeled with fluorescein at the 5'-end.

In three separate experiments, the wild-type target, the mutant target, and an equimolar mixture of both targets was exposed (0.1 nM wt508, 0.1 nM mu508, and 0.1 nM wt508 plus 0.1 nM mu508, respectively, in a solution compatible with nucleic acid hybridization) to a CF chip. The hybridization mixture was incubated overnight at room temperature, and then the chip was scanned on a reader (a confocal fluorescence microscope in photon-counting mode); images of the chip were constructed from the photon counts) at several successively higher temperatures while still in contact with the target solution. After each temperature change, the chip was allowed to equilibrate for approximately one-half hour before being scanned. After each set of scans, the chip was exposed to denaturing solvent and conditions to wash, *i.e.*, remove target that had bound, the chip so that the next experiment could be done with a clean chip.

The results of the experiments are shown in Figures 18, 19, 20, and 21. Figure 18, in panels A, B, and C, shows an image made from the region of a DNA chip containing CFTR exon 10 probes; in panel A, the chip was hybridized to a wild-type target; in panel C, the chip was hybridized to a mutant ΔF508 target; and in panel B, the chip was hybridized to a mixture

of the wild-type and mutant targets. Figure 19, in sheets 1 - 3, corresponding to panels A, B, and C of Figure 3, shows graphs of fluorescence intensity versus tiling position. The labels on the horizontal axis show the bases in the wild-type sequence corresponding to the position of substitution in the respective probes. Plotted are the intensities observed from the features (or synthesis sites) containing wild-type probes, the features containing the substitution probes that bound the most target ("called"), and the feature containing the substitution probes that bound the target with the second highest intensity of all the substitution probes ("2nd Highest").

These figures show that, for the wild-type target and the equimolar mixture of targets, the substitution probe with a nucleotide sequence identical to the corresponding wild-type probe bound the most target, allowing for an unambiguous assignment of target sequence as shown by letters near the points on the curve. The target wt508 thus hybridized to the probes in the wild-type lane of the chip, although the strength of the hybridization varied from probe-to-probe, probably due to differences in melting temperature. The sequence of most of the target can thus be read directly from the chip, by inference from the pattern of hybridization in the lanes of substitution probes (if the target hybridizes most intensely to the probe in the A-lane, then one infers that the target has a T in the position of substitution, and so on).

For the mutant target, the sequence could similarly be called on the 3'-side of the deletion. However, the intensity of binding declined precipitously as the point of substitution approached the site of the deletion from the 3'-end of the target, so that the binding intensity on the wild-type probe whose point of substitution corresponds to the T at the 3'-end of the deletion was very close to background. Following that pattern, the wild-type probe whose point of substitution corresponds to the middle base (also a T) of the deletion bound still less target. However, the probe in the T-lane of that column set bound the target very well. Examination of

the sequences of the two targets reveals that the deletion places an A at that position when the sequences are aligned at their 3'-ends and that the T-lane probe is complementary to the mutant target with but two mismatches near an end (shown below in lower-case letters, with the position of substitution underlined):

Target: 5'-CATTAAAGAAAATATCATTGGTGTTTCCTATGATGA

Probe: 3'-TagTAGTAACCACAA

Thus the T-lane probe in that column set calls the correct base from the mutant sequence. Note that, in the graph for the equimolar mixture of the two targets, that T-lane probe binds almost as much target as does the A-lane probe in the same column set, whereas in the other column sets, the probes that do not have wild-type sequence do not bind target at all as well. Thus, that one column set, and in particular the T-lane probe within that set, detects the ΔF508 mutation under conditions that simulate the homozygous case and also conditions that simulate the heterozygous case.

Although in this example the sequence could not be reliably deduced near the ends of the target, where there is not enough overlap between target and probe to allow effective hybridization, and around the center of the target, where hybridization was weak for some other reason, perhaps high AT-content, the results show the method and the probes of the invention can be used to detect the mutation of interest. The mutant target gave a pattern of hybridization that was very similar to that of the wt508 target at the ends, where the two share a common sequence, and very different in the middle, where the deletion is located. As one scans the image from right to left, the intensity of hybridization of the target to the probes in the wild-type lane drops off much more rapidly near the center of the image for mu508 than for wt508; in addition, there is one probe in the T-lane that hybridizes intensely with mu508 and hardly at all with wt508. The results from the equimolar mixture of the two targets, which represents the case one would encounter in testing a heterozygous individual for the mutation, are a blend of the results for the separate targets, showing the power of the

invention to distinguish a wild-type target sequence from one containing the ΔF508 mutation and to detect a mixture of the two sequences.

The results above clearly demonstrate how the DNA chips of the invention can be used to detect a deletion mutation, ΔF508; another model system was used to show that the chips can also be used to detect a point mutation as well. One mutation in the CFTR gene is G480C, which involves the replacement of the G in position 46 of exon 10 by a T, resulting in the substitution of a cysteine for the glycine normally in position #480 of the CFTR protein. The model target sequences included the 21-mer probe wt480 to represent the wild-type sequence at positions 37-55 of exon 10: 5'-CCTTCAGAGGGTAAAATTAAG and the 21-mer probe mu480 to represent the mutant sequence: 5'-CCTTCAGAGTGTAAAATTAAG.

In separate experiments, a DNA chip was hybridized to each of the targets wt480 and mu480, respectively, and then scanned with a confocal microscope. Figure 20, in panels A, B, and C, shows an image made from the region of a DNA chip containing CFTR exon 10 probes; in panel A, the chip was hybridized to the wt480 target; in panel C, the chip was hybridized to the mu480 target; and in panel B, the chip was hybridized to a mixture of the wild-type and mutant targets. Figure 21, in sheets 1 - 3, corresponding to panels A, B, and C of Figure 20, shows graphs of fluorescence intensity versus tiling position. The labels on the horizontal axis show the bases in the wild-type sequence corresponding to the position of substitution in the respective probes. Plotted are the intensities observed from the features (or synthesis sites) containing wild-type probes, the features containing the substitution probes that bound the most target ("called"), and the feature containing the substitution probes that bound the target with the second highest intensity of all the substitution probes ("2nd Highest").

These figures show that the chip could be used to sequence a 16-base stretch from the center of the target wt480 and that discrimination against mismatches is quite good

throughout the sequenced region. When the DNA chip was exposed to the target mu480, only one probe in the portion of the chip shown bound the target well: the probe in the set of probes devoted to identifying the base at position 46 in exon 10 and that has an A in the position of substitution and so is fully complementary to the central portion of the mutant target. All other probes in that region of the chip have at least one mismatch with the mutant target and therefore bind much less of it. In spite of that fact, the sequence of mu480 for several positions to both sides of the mutation can be read from the chip, albeit with much-reduced intensities from those observed with the wild-type target.

The results also show that, when the two targets were mixed together and exposed to the chip, the hybridization pattern observed was a combination of the other two patterns. The wild-type sequence could easily be read from the chip, but the probe that bound the mu480 target so well when only the mu480 target was present also bound it well when both the mutant and wild-type targets were present in a mixture, making the hybridization pattern easily distinguishable from that of the wild-type target alone. These results again show the power of the DNA chips of the invention to detect point mutations in both homo- and heterozygous individuals.

To demonstrate clinical application of the DNA chips of the invention, the chips were used to study and detect mutations in nucleic acids from genomic samples. Genomic samples from a individual carrying only the wild-type gene and an individual heterozygous for ΔF508 were amplified by PCR using exon 10 primers containing the promoter for T7 RNA polymerase. Illustrative primers of the invention are shown below.

Exon	Name	Sequence
10	CFi9-T7	TAATACGACTCACTATAGGGAGatgacctaataatgatgggttt
10	CFi10c-T7	TAATACGACTCACTATAGGGAGtagtgtgaagggttcatatgc
10	CFi10c-T3	CTCGGAATTAACCCTCACTAAAGGtagtgtgaagggttcatatgc
11	CFi10-T7	TAATACGACTCACTATAGGGAGagcatactaaaagtgactctc
11	CFi11c-T7	TAATACGACTCACTATAGGGAGacatgaatgacatttacagcaa
11	CFi11c-T3	CGGAATTAACCCTCACTAAAGGacatgaatgacatttacagcaa

These primers can be used to amplify exon 10 or exon 11 sequences; in another embodiment,. multiplex PCR is employed, using two or more pairs of primers to amplify more than one exon at a time.

The product of amplification was then used as a template for the RNA polymerase, with fluoresceinated UTP present to label the RNA product. After sufficient RNA was made, it was fragmented and applied to an exon 10 DNA chip for 15 minutes, after which the chip was washed with hybridization buffer and scanned with the fluorescence microscope. A useful positive control included on many CF exon 10 chips is the 8-mer 3'-CGCCGCCG-5'. Figure 22, in panels A and B, shows an image made from a region of a DNA chip containing CFTR exon 10 probes; in panel A, the chip was hybridized to nucleic acid derived from the genomic DNA of an individual with wild-type ΔF508 sequences; in panel B, the target nucleic acid originated from a heterozygous (with respect to the ΔF508 mutation) individual. Figure 23, in sheets 1 and 2, corresponding to panels A and B of Figure 22, shows graphs of fluorescence intensity versus tiling position.

These figures show that the sequence of the wild-type RNA can be called for most of the bases near the mutation. In the case of the ΔF508 heterozygous carrier, one particular probe, the same one that distinguished so clearly between the wild-type and mutant oligonucleotide targets in the model system described above, in the T-lane binds a large amount of RNA, while the same probe binds little RNA from the wild-type individual. These results show that the DNA chips of the invention are capable of detecting the ΔF508 mutation in a heterozygous carrier.

Further chips were constructed using the block tiling strategy to provide an array of probes for analyzing a CFTR mutation. The array comprised 93 mm x 96 μm features arranged into eleven columns and four rows (44 total probes). Probes in five of these columns were from four probe sets tiled based on the wildtype CFTR sequence and having interrogation positions corresponding to the site of a mutation and two bases on either side. Five of the remaining columns contained

four sets of probes tiled based on the mutant version of the CFTR sequence. These probe sets also had interrogation positions corresponding to the site of mutation and two nucleotides on either side. The eleventh column contained four cells for control probes.

Fluorescently labeled hybridization targets were prepared by PCR amplification. 100 μg of genomic DNA, 0.4 μM of each primer, 50 μM each dATP, dCTP, dCTP and dUTP (Pharmacia) n 10mM Tris-Cl, pH 8.3, 50 mM KCl, 2.5 mM $MgCl_2$ and 2 U Taq polymerase (Perkin-Elmer) were cycled 36 times using a Perkin-Elmer 9600 thermocycler and the following times and temperatures: 95°C, 10 sec., 55°C, 10 sec., 72°C, 30 sec. 10 μl of this reaction product was used as a template in a second, asymmetric PCR reaction. Conditions included 1μM asymmetric PCR primer, 50 μM each dATP, dCTP, TTP, 25 μM fluorescein-dGTP (DuPont), 10 mM Tris-Cl, pH 9.1, 75 mM KCl, 3.5 mM $MgCl_2$. The reaction was cycled 5X with the following conditions: 95°C, 10 sec, 60°C, 10 sec, 55°C, 1 min. and 72°C, 1.5 min. This was immediately followed with another 20 cycles using the following conditions: 95°C, 10 sec, 60°C, 10 sec., 72°C, 1.5 min.

Amplification products were fragmented by treating with 2 U of Uracil-N-glycosylase (Gibco) at 30°C for 30 min. followed by heat denaturation at 95°C for 5 min. Finally, the labeled, fragmented PCR product was diluted into hybridization buffer made up of 5 X SSPE and 1 mM Cetyltrimethylammonium Bromide (CTAB). The dilution factor ranged from 10x to 25x with 40 μl of sample being diluted into 0.4 ml to 1 ml of hybridization solution.

Target hybridization was generally carried out with the chip shaking in a small dish containing 500 μl to 1 ml total volume of hybridization solution. All hybridizations were done at 30°C constant temperature. Alternatively, some hybridizations were carried out with chips enclosed in a plastic package with the 1 cm x 1 cm chip glued facing a 250 μl fluid chamber. 250-350 μl of hybridization solution was introduced and mixed using a syringe pump. Temperature was controlled by interfacing the back surface of the package with

a Peltier heating/cooling device. Following hybridization chips were washed with 5X SSPE, 0.1% Triton X-100 at 25°C-30°C prior to fluorescent image generation.

Hybridized, washed DNA chips were scanned for fluorescence using a stage-scanning confocal epifluorescent microscope and 488nm argon ion laser excitation. Emitted light was collected through a band pass filter centered at 530nM. The resulting fluorescence image was spatially reconstructed and intensity data were then analyzed. Features with the peak fluorescence intensity in each column were identified and compared with any signal intensity at the remaining single base mismatch probe sites in the same column. The sequences of the highest intensity features were then compared across all ten columns of each sub-array to determine whether peak intensity scores for the wild type sequence and the mutant sequence were similar or significantly different. These results were used to generate the genotype call of wild type (high intensity signals only in wild type probe columns), mutant (high intensity signals only in the mutant probe columns) or heterozygous (high intensity signals in both the wild type and mutant probe columns).

Figure 24 (panel A) shows an image of the fluorescence signals in arrays designed to detect the G551D(G>A) and Q552X(C>T) CFTR mutations. The hybridization target is an exon 11 amplicon generated from wild type genomic DNA. Wild type hybridization patterns are evident at both locations. No significant fluorescence signal resulted at any of the features with probes complementary to mutant or mismatched sequences. Relative fluorescence intensities were six fold brighter for the perfect matched wildtype features compared with the background signal intensity at mutant and mismatch features. In addition, the sequence at these loci can be confirmed as AGGTC and GTCAA, respectively, where the bold type face indicates the mutation sites. Figure 24 (panel B) shows the same probe array features after hybridization with a fluorescent target generated from DNA heterozygous for the G551D mutation. Both the wild type and mutant probe columns have features with significant fluorescence intensity,

indicating the hybridization of both wild type and mutant CFTR alleles at this site. Only wildtype probes hybridized with any significant fluorescence signal in the Q552X subarray indicating a wild type target sequence. However, an additional feature that did not hybridize in the first experiment shows significant fluorescence intensity in this experiment. Because the G551D and Q552X mutations are only two bases apart, the a probe sequence in the additional feature has a perfectly matched 12-mer overlap with the mutant G551D target.

Figure 25 (panels A and B) illustrates mutation analysis for ΔF508, a three base pair deletion in Exon 10 of the CFTR gene. In contrast to the hybridization pattern seen in base change mutations, in mutations where bases are inserted or deleted, probe arrays show a different hybridization pattern. Identical probes are synthesized in the two central columns of base substitution arrays. As a result, either mutant or wild type target hybridizations always result in two side-by-side features (a doublet) with high fluorescence intensity at the center of the array. In a heterozygote hybridization, two sets of doublets, one matched to the wild type sequence and one to the mutant sequence occur (Figure 24, panel B). In contrast, wild type and mutant probe column sequences are offset from each other for deletion or insertion mutations and hybridization doublets are not seen. Instead of the six high intensity signals with one doublet, five independent features in alternating columns characterize a homozygote and ten features, one in each column will be positive with heterozygote targets. This is evident from the ΔF508 hybridization pattern in Figure 25, panel A. Although a wildtype target has been hybridized and the highest intensity features confirm the wild type sequence (ATCTT), there is an additional hybridization in the first mutant column. Analysis of that probe sequence shows a 10 base perfect match with the mutant sequence.

The image in Figure 25, panel B resulted from hybridizing a DNA chip with a target homozygous for ΔF508. In this image five features, all with probe sequences

complementary to the mutant show significant signal. The mutation sequence bridging the deletion site, ATTGG, is confirmed. Similar to what was seen in the example of the G551D mutation, there is added information in neighboring subarrays designed to detect the ΔI507 and F508C mutations. This is expected since they are in such close proximity to ΔF508 that their probe sets significantly overlap the ΔF508 probes. The ΔF508 homozygous target has no perfect matches with wild type or mutant probes in the ΔI507 and F508C subarrays. However, there are some low intensity signals within these two blocks of probes. The F508C array has a doublet that matches 11 bases of the mutant ΔF508 target. Similarly, the hybridization in the eighth column of the ΔI507 array has a probe that matches 13/14 bases with the target.

Figure 26 shows hybridization of a heterozygous double mutant ΔF508/F508C to the same array as described above. Conventional reverse dot blot would score this sample as a homozygous ΔF508 mutant. In the present assays, the ΔF508 and F508C alleles are separately detected by the respective subarrays designed to detect these mutations.

C. Chips for Cancer Diagnosis

There are at least two types of genes which are often altered in cancerous cells. The first type of gene is an oncogene such as a mismatch-repair gene, and the second type of gene is a tumor suppressor gene such as a transcription factor. Examples of mismatch repair oncogenes genes include hMSH2 (Fishel et al., *Cell* 75, 1027-1038 (1993)) and hMLH1 (Papadopoulos et al., *Science* 263, 1625-1628 (1994)). The most well-known example of a tumor suppressor gene is the p53 protein gene (Buchman et al., *Gene* 70, 245-252 (1988). By monitoring the state of both oncogenes and tumor suppressor genes (individually and in combination) in a patient, it is possible to determine individual susceptibility to a cancer, a patient's prognosis upon cancer diagnosis, and to target therapy more efficiently.

The p53 gene spans 20 kbp in humans and has 11 exons, 10 of which are protein coding (*see* Tominaga et al., 1992,

Critical Reviews in Oncogenesis 3:257-282, incorporated herein by reference). The gene produces a 53 kilodalton phosphoprotein that regulates DNA replication. The protein acts to halt replication at the G1/S boundary in the cell cycle and is believed to act as a "molecular policeman," shutting down replication when the DNA is damaged or blocking the reproduction of DNA viruses (*see* Lane, 1992, *Nature* 358:15-16, incorporated herein by reference). The p53 transcription factor is part of a fundamental pathway which controls cell growth. Wild-type p53 can halt cell growth, or in some cases bring about programmed cell death (apoptosis). Such tumor-suppressive effects are absent in a variety of known p53 gene mutations. Moreover, p53 mutants not only deprive a cell of wild-type p53 tumor suppression, they also may spur abnormal cell growth.

In tumor cells, p53 is the most commonly mutated gene discovered to date (*see* Levine et al., 1991, *Nature* 351:453-456, and Hollstein et al., 1991, *Science* 253:49-53, each of which is incorporated herein by reference) Over half of the 6.5 million patients diagnosed with cancer annually possess p53 mutations in their tumor cells. Among common tumors, about 70% of colorectal cancers, 50% of lung cancers and 40% of breast cancers contain p53 mutations. In all, over 51 types of human tumors have been documented to possess p53 mutations, including bladder, brain, breast, cervix, colon, esophagus, larynx, liver, lung, ovary, pancreas, prostate, skin, stomach, and thyroid tumors (Culotta & Koshland, *Science* 262, 1958-1961 (1993); Rodrigues et al., 1990, *PNAS* 87:7555-7559, incorporated herein by reference). According to data presented by David Sidransky (1992 San Diego Conference), over 400 mutations in p53 are known. The presence of a p53 mutation in a tumor has also been correlated with a patient's prognosis. Patients who possess p53 mutations have a lower 5-year survival rate.

Proper diagnosis of the form of p53 in tumor cells is critical to clinicians to prescribe appropriate therapeutic regimens. For instance, patients with breast cancer who show no invasion of nearby lymph nodes generally do not relapse

after standard surgical treatment and chemotherapy. Of the 25% who do relapse after surgery and chemotherapy, additional chemotherapy is appropriate. At present, there is no clear way to determine which patients will benefit from such additional chemotherapy prior to relapse. However, correlating p53 mutations to tumorigenicity and metastasis provides clinicians with a means to determine whether such additional treatments are warranted.

In addition to facilitating conventional chemotherapy, appropriate diagnosis of p53 mutations provides clinicians with the ability to identify individuals who will benefit the most from gene therapy techniques, in which appropriately operative p53 copies are restored to a tumor site. Clinical p53 gene therapy trials are presently underway (Culotta & Koshland, *supra*).

The analysis of p53 mutations can also be used to identify which carcinogens lead to particular tumors (Harris, *Science* 262, 1980-1981 (1993)). For instance, dietary aflatoxin B_1 exposure is associated with G:C to T:A transversions at residue 249 of p53 in hepatocellular carcinomas (Hsu et al., *Nature* 350, 427 (1991); Bressac et al., *Nature* 350, 429 (1991); Harris, *supra*).

While most described p53 mutations are somatic in origin, some types of cancer are associated with germline p53 mutation. For instance, Li-Fraumeni syndrome is a hereditary condition in which individuals receive mutant p53 alleles, resulting in the early onset of various cancers (Harris, *supra*); Frebourg et al., *PNAS* 89, 6413-6417 (1992); Malkin et al., *Science* 250, 1233 (1990)). These mutations are associated with instability in the rest of the genome, creating multiple genetic alterations, and eventually leading to cancer.

hMLH1 and hMSH2 are mismatch repair genes which are causal agents in hereditary nonpolyposis colorectal cancer in individuals with mutant hMLH1 or hMSH2 alleles (Fishel et al., *supra*, and Papadopoulos et al., *supra*). Hereditary nonpolyposis colorectal cancer is a common genetic disorders, affecting about 1 in 200 individuals (Lynch et al.,

Gastroenterology 104, 1535 (1993)). Detection of hMLH1 and hMSH2 mutations in the population allows diagnosis of nonpolyposis colorectal cancer prone individuals prior to the manifestation of disease. This allows for the implementation of special screening programs for cancer-prone individuals to ensure early detection of cancer, thereby enhancing survival rates of afflicted individuals. In addition, genetic counselors may use the information derived from HMLH1 and HMSH2 chips to improve family planning as described for cystic fibrosis chips. The detection of mutations in hMLH1 and hMSH2 individually or in combination with p53 can also be used by clinicians to assess cancer prognosis and treatment modality. Finally, the information can be used to target appropriate individuals for gene therapy.

The entire hMLH1 gene is less than 85 kbp in length, comprising 2268 coding nucleotides (Papadopoulos et al., *supra*). Sequences from the gene have been deposited with GenBank (accession number U07418). Mutations associated with hereditary nonpolyposis colorectal cancer include the deletion of exon 5 (codons 578-632), a 4 base pair deletion of codons 727 and 728 resulting in a shift in the reading frame of the gene, a 4 base pair insertion at codons 755 and 756 resulting in an extension of the COOH terminus, a 371 base pair deletion and frameshift mutation at position 347, and a transversion causing an alteration of codon 252 resulting in the insertion of a stop codon (*id.*).

hMSH2 is a human homologue of the bacterial *MutS* and *S. cerevisiae* MSH mismatch-repair genes. MSH2, like hMLH1 is associated with hereditary nonpolyposis cancer. Although only a few MSH2 gene samples from tumor tissue have been characterized, at least some tumor samples show a T to C transition mutation at position 2020 of the cDNA sequence, resulting in the loss of an intron-exon splice acceptor site.

In view of the role of mutations in p53, MSH2 and/or hMLH1 in hereditary predisposition to cancer, to neoplastic transformation events leading to cancer and to cancer prognosis, it is important to screen individuals to determine whether they possess mutant alleles, and to identify precisely

which mutations the individuals possess. Because many mutations are point mutations, or extremely small insertions or deletions, which are generally undetectable by standard Southern analysis, accurate diagnosis requires a capacity to examine a gene nucleotide-by-nucleotide.

Mutations in the hMSH2, hMLH1 or p53 genes, irrespective of whether previously characterized, can be detected by any of the tiling strategies noted above. Reference sequences of interest include full-length genomic and cDNA sequences of each of these genes and subsequences thereof, such as exons and introns. For example, each nucleotide in the 20 kb p53 genomic sequence can be tiled using the basic strategy with an array of about 80,000 probes. As in the CFTR chip, some reference sequences are comparatively short sequences including the site of a known mutation and a few flanking nucleotides. Some chips tile reference sequences that encompass mutational "hot spots." For instance, a variety of cellular and oncoviral proteins bind to specific regions of p53, including Mdm2, SV40 T antigen, E1b from adenovirus and E6 from human papilloma virus. These binding sites correlate to some extent with observed high frequency somatic mutation regions of p53 found in tumor cells from cancer patients (*see* Harris et al., *supra*). Hot spots include exons 2, 3, 5, 6, 7 and 8 and the intronic regions between exons 2 and 3, 3 and 4 and 4 and 5. Fragments of the hMLH1 gene of particular interest include those encoding codons 578-632, 727, 728, 347, 252. Some chips are tiled to read mutations in each of the hMSH2, hMLH1 and p53 genes, both wildtype and mutant versions.

Standard or asymmetric PCR can be used to generate the target DNA used in the tiling assays described above. In general, PCR is used to amplify hMSH2, hMLH1 or p53 sequences from a tissue of interest such as a tumor. Mixed PCR reactions can also be used to generate hMSH2, hMLH1 or p53 sequences simultaneously in a single reaction mixture. Any of the coding or noncoding sequences from the genes may be amplified for use in the block tiling assays described above.

Table 8 below provides examples of primers which are useful in synthesizing specific regions of hMSH2, hMHLH1 and

p53. Other primers can readily be devised from the known genomic and cDNA sequences of the genes. The primers described in Table 8 specific for p53 amplification have ends tailored to facilitate cloning into standard restriction enzyme cloning sites.

Table 8: Examples of PCR primers useful in amplifying regions of p53, hMHH1 and hMSH2.

Region Amplified	Primer Sequence	Description
Exon 5 (p53)	TAA TAC GAC TCA CTA TAG GGA GA CCC TGG GCA ACC AGC CCT GTC GT	Exon 5 T7 Primer (5' T7 to p53 3').
Exon 5 (p53)	ATG CAA TTA ACC CTC ACT AAA GGG AGA CAC TTG TGC CCT GAC TTT CAA C	Exon 5 T3 Primer (5' T3 to p53 3').
Exon 6 (p53)	TAA TAC GAC TCA CTA TAG GGA GCC TCC TCC CAG AGA CCC	Exon 6 T7 Primer (5'T7 to p53 3').
Exon 6 (p53)	ATG CAA TTA ACC CTC ACT AA GGG AGA TCC CCA GGC CTC TGA TTC CTC ACT G	Exon 6 T3 Primer (5'T3 to p53 3').
Exon 7 (p53)	TAA TAC GAC TCA CTA TAG GGA CTG GGG CAC AGC CAG GCC AGT GTG CA	Exon 7 T7 Primer (5' T7 to p53 3').
Exon 7 (p53)	ATG CAA TTA ACC CTC ACT AAA GGG AGA GTC TCC CCA AGG CGC ACT GGC CTC A	Exon 7 T3 Primer (5' T3 to p53 3').
Exon 8 (p53)	TAA TAC GAC TCA CTA TAG GGA GGG CAT AAC TGC ACC CTT GGT CTC CTC C	Exon 8 T7 Primer (5' T7 to p53 3').
Exon 8 (p53)	ATG CAA TTA ACC CTC ACT AAA GGG AGA GGA CCT GAT TTC CTT ACT GCC TCT TGC	Exon 8 T3 Primer (5' T3 to p53 3').
hMSH2	GAC ATG GCG GTG CAG CCG AAG GAG A	Primer for MSH2, 5' to 3'. If used with MSH2 primer below, a 3033 base pair amplicon will result
hMSH2	CTA TGT CAA TTG CAA ACA GTG CTC AGT TAC AG	Primer for hMSH2 5'to 3'.
hMLH1	CTT GGC TCT TCT GGC GCC AAA ATG TCG TTC	Primer for hMLH1, 5'to 3'. If used with hMLH1 primer below, a 2484 base pair amplicon will result.
hMLH1	TAT GTT AAG ACA CAT CTA TTT ATT TAT AAT CAA TCC	Primer for hMLH1 5' to 3'.

After PCR amplification of the target amplicon one strand of the amplicon can be isolated, *i.e.*, using a biotinylated primer that allows capture of the undesired strand on streptavidin beads. Alternatively, asymmetric PCR can be used to generate a single-stranded target. Another approach involves the generation of single stranded RNA from the PCR product by incorporating a T7 or other RNA polymerase promoter in one of the primers. The single-stranded material can optionally be fragmented to generate smaller nucleic acids with less significant secondary structure than longer nucleic acids.

In one such method, fragmentation is combined with labeling. To illustrate, degenerate 8-mers or other degenerate short oligonucleotides are hybridized to the single-stranded target material. In the next step, a DNA polymerase is added with the four different dideoxynucleotides, each labeled with a different fluorophore. Fluorophore-labeled dideoxynucleotide are available from a variety of commercial suppliers. Hybridized 8-mers are extended by a labeled dideoxynucleotide. After an optional purification step, *i.e.*, with a size exclusion column, the labeled 9-mers are hybridized to the chip. Other methods of target fragmentation can be employed. The single-stranded DNA can be fragmented by partial degradation with a DNAse or partial depurination with acid. Labeling can be accomplished in a separate step, *i.e.*, fluorophore-labeled nucleotides are incorporated before the fragmentation step or a DNA binding fluorophore, such as ethidium homodimer, is attached to the target after fragmentation.

Exemplary Chips

a. Exon VI Chip

To illustrate the value of the DNA chips of the present invention in such a method, a DNA chip was synthesized by the VLSIPS™ method to provide an array of overlapping probes which represent or tile across a 60 base region of exon 6 of the p53 gene. To demonstrate the ability to detect substitution mutations in the target, twelve different single substitution

mutations (wild type and three different substitutions at each of three positions) were represented on the chip along with the wild type. Each of these mutations was represented by a series of twelve 12-mer oligonucleotide probes, which were complementary to the wild type target except at the one substituted base. Each of the twelve probes was complementary to a different region of the target and contained the mutated base at a different position, e.g., if the substitution was at base 32, the set of probes would be complementary--with the exception of base 32--to regions of the target 21-32, 22-33, and 32-43). This enabled investigation of the effect of the substitution position within the probe. The alignment of some of the probes with a 12-mer model target nucleic acid is shown in Figure 27.

To demonstrate the effect of probe length, an additional series of ten 10-mer probes was included for each mutation (see Figure 28). In the vicinity of the substituted positions, the wild-type sequence was represented by every possible overlapping 12-mer and 10-mer probe. To simplify comparisons, the probes corresponding to each varied position were arranged on the chip in the rectangular regions with the following structure: each row of cells represents one substitution, with the top row representing the wild type. Each column contains probes complementary to the same region of the target, with probes complementary to the 3'-end of the target on the left and probes complementary to the 5'-end of the target on the right. The difference between two adjacent columns is a single base shift in the positioning of the probes. Whenever possible, the series of 10-mer probes were placed in four rows immediately underneath and aligned with the 4 rows of 12-mer probes for the same mutation.

To provide model targets, 5' fluoresceinated 12-mers containing all possible substitutions in the first position of codon 192 were synthesized (see the starred position in the target in Figure 27). Solutions containing 10 nM target DNA in 6X SSPE, 0.25% Triton X-100 were hybridized to the chip at room temperature for several hours. While target nucleic was hybridized to the chip, the fluorophores on the chip were

excited by light from an argon laser, and the chip was scanned with an autofocusing confocal microscope. The emitted signals were processed by a PC to produce an image using image analysis software. By 1 to 3 hours, the signal had reached a plateau; to remove the hybridized target and allow hybridization to another target, the chip was stripped with 60% formamide, 2 X SSPE at 17 °C for 5 minutes. The washing buffer and temperature can vary, but the buffer typically contains 2-to-3X SSPE, 10-to-60% formamide (one can use multiple washes, increasing the formamide concentration by 10% each wash, and scanning between washes to determine when the wash is complete), and optionally a small percentage of Triton X-100, and the temperature is typically in the range of 15-to-18°C

Very distinct patterns were observed after hybridization with targets with 1 base substitutions and visualization with a confocal microscope and software analysis, as shown in Figure 29. In general, the probes which form perfect matches with the target retain the highest signal. For example, in the first image, the 12-mer probes that form perfect matches with the wild-type (WT) target are in the first row (top). The 12-mer probes with single base mismatches are located in the second, third, and fourth rows and have much lower signals. The data is also depicted graphically in Figure 30. On each graph, the X ordinate is the position of the probe in its row on the chip, and the Y ordinate is the signal at that probe site after hybridization. When a target with a different one base substitution is hybridized the complementary set of probes has the highest signal (see pictures 2, 3, and 4 in Figure 29 and graphs 2, 3, and 4 in Figure 30). In each case, the probe set with no mismatches with the target has the highest signals. Within a 12-mer probe set, the signal was highest at position 6 or 7. The graphs show that the signal difference between 12-mer probes at the same X ordinate tended to be greatest at positions 5 and 8 when the target and the complementary probes formed 10 base pairs and 11 base pairs, respectively. Because tumors often have both WT and mutant p53 genes, mixed target

populations were also hybridized to the chip, as shown in Figure 31. When the hybridization solution consisted of a 1:1 mixture of WT 12-mer and a 12-mer with a substitution in position 7 of the target, the sets of probes that were perfectly matched to both targets showed higher signals than the other probe sets.

The hybridization efficiency of a 10-mer probe array as compared to a 12-mer probe array was also compared. The 10-mer and 12-mer probe arrays gave comparable signals (see graphs 1-4 in Figure 30 and graphs 1-4 in Figure 32). However, the 10-mer probe sets, which are in rows 5-8 (see images in Figure 29), seemed to be better in this model system than the 12-mer probe sets at resolving one target from another, consistent with the expectation that one base mismatches are more destabilizing for 10-mers than 12-mers. Hybridization results within probe sets perfectly matched to target also followed the expectation that, the more matches the individual probe formed with the target, the higher the signal. However, duplexes with two 3' dangles (see Figure 30, position 6 in graphs 1-4) have about as much signal as the probes which are matched along their entire length (see Figure 30, position 7, in graphs 1-4).

This illustrative model system shows that 12-mer targets that differ by one base substitutions can be readily distinguished from one another by the novel probe array provided by the invention and that resolution of the different 12-mer targets was somewhat better with the 10-mer probe sets than with the 12-mer probe sets.

b. Exon V Chip

To analyze DNA from exon 5 of the p53 tumor suppressor gene, a set of overlapping 17-mer probes was synthesized on a chip. The probes for the WT allele were synthesized so as to tile across the entire exon with single base overlaps between probes. For each WT probe, a sets of 4 additional probes, one for each possible base substitution at position 7, were synthesized and placed in a column relative to the WT probe. Exon 5 DNA was amplified by PCR with primers flanking the exon. One of the primers was labeled with fluorescein; the

other primer was labeled with biotin. After amplification, the biotinylated strand was removed by binding to streptavidin beads. The fluoresceinated strand was used in hybridization.

About 1/3 of the amplified, single-stranded nucleic acid was hybridized overnight in 5 X SSPE at 60°C to the probe chip (under a cover slip). After washing with 6 X SSPE, the chip was scanned using confocal microscopy. Figure 33 shows an image of the p53 chip hybridized to the target DNA. Analysis of the intensity data showed that 93.5% of the 184 bases of exon 5 were called in agreement with the WT sequence (see Buchman et al., 1988, Gene 70: 245-252, incorporated herein by reference). The miscalled bases were from positions where probe signal intensities were tied (1.6%) and where non-WT probes had the highest signal intensity (4.9%). Figure 34 illustrates how the actual sequence was read. Gaps in the sequence of letters in the WT rows correspond to control probes or sites. Positions at which bases are miscalled are represented by letters in italic type in cells corresponding to probes in which the WT bases have been substituted by other bases.

As the diagram indicates, the miscalled bases are from the low intensity areas of the image, which may be due to secondary structure in the target or probes preventing intermolecular hybridization. To diminish the effects due to secondary structure, one can employ shorter targets (i.e., by target fragmentation) or use more stringent hybridization conditions. In addition, the use of a set of probes synthesized by tiling across the other strand of a duplex target can also provide sequence information buried in secondary structure in the other strand. It should be appreciated, however, that the pattern of low intensity areas that forms as a result of secondary structure in the target itself provides a means to identify that a specific target sequence is present in a sample. Other factors that may contribute to lower signal intensities include differences in probe densities and hybridization stabilities.

These results demonstrate the advantages provided by the DNA chips of the invention to genetic analysis. As another example, heterozygous mutations are currently sequenced by an arduous process involving cloning and repurification of DNA. The cloning step is required, because the gel sequencing systems are poor at resolving even a 1:1 mixture of DNA. First, the target DNA is amplified by PCR with primers allowing easy ligation into a vector, which is taken up by transformation of E. coli, which in turn must be cultured, typically on plates overnight. After growth of the bacteria, DNA is purified in a procedure that typically takes about 2 hours; then, the sequencing reactions are performed, which takes at least another hour, and the samples are run on the gel for several hours, the duration depending on the length of the fragment to be sequenced. By contrast, the present invention provides direct analysis of the PCR amplified material after brief transcription and fragmentation steps, saving days of time and labor.

D. Mitochondrial Genome Chips

A human cell may have several hundred mitochondria, each with more than one copy of mtDNA. There is strand asymmetry in the base compositions, with one strand (Heavy) being relatively G rich, and the other strand (Light) being C rich. The L strand is 30.9% A, 31.2% C, 13.1% G, and 24.7% T. Human mtDNA is information-rich, encoding some 22 tRNAs, 12S and 16S rRNAs, and 13 polypeptides involved in oxidative phosphorylation. No introns have been detected. RNAs are processed by cleavage at tRNA sequences, and polyadenylated postranscriptionally. In some transcripts, polyadenylation also creates the stop codon, illustrating the parsimony of coding. In many individuals, mtDNA can be treated as haploid. However, some individuals are heteroplasmic (have more than one mtDNA sequence), and the degree of heteroplasmy can vary from tissue to tissue. Also, the rate of replication of mtDNAs can differ and together with random segregation during cell division, can lead to changes in heteroplasmy over time.

The human mitochondrial genome is 16,569 nucleotides

long. The sequence of the L-strand is numbered arbitrarily from the MboI-5/7 boundary in the D-loop region. The complete sequence of the human mitochondrial genome has been published. *See* Anderson et al., *Nature* 290, 457-465 (1981).

Mitochondrial DNA is maternally inherited, and has a mutation rate estimated to be tenfold higher than single copy nuclear DNA (Brown et al., *Proc. Natl. Acad. Sci. USA* 76, 1967-1971 (1979)). Human mtDNAs differ, on average, by about 70 base substitutions (Wallace, *Ann. Rev. Biochem.* 61, 1175-1212 (1992)). Over 80% of substitutions are transitions (*i.e.*, pyrimidine-pyrimidine or purine-purine).

Analysis of mitochondrial DNA serves several purposes. Detection of mutations in the mitochondrial genome allows diagnosis of a number of diseases. The mitochondrial genome has been identified as the locus of several mutations associated with human diseases. Some of the mutations result in stop codons in structural genes. Such mutations have been mapped and associated with diseases, such as Leber's hereditary optic neuropathy, neurogenic muscular weakness, ataxia and retinitis pigmentosa. Other mutations (nucleotide substitutions) occur in tRNA coding sequences, and presumably cause conformational defects in transcribed tRNA molecules. Such mutations have also been mapped and associated with diseases such as Myoclonic Epilepsy and Ragged Red Fiber Disease. Another type of mutation commonly found is deletions and/or insertions. Some deletions span segments of several kb. Again, such mutations have been mapped and associated with diseases, for example, ocular myopathy and Person Syndrome. *See* Wallace, *Ann. Rev. Biochem.* 61-1175-1212 (1992) (incorporated by reference in its entirety for all purposes). Early detection of such diseases allows metabolic or genetic therapy to be administered before irretrievable damage has occurred. *Id.* Analysis of mitochondrial DNA is also important for forensic screening. Because the mitochondrial genome is a locus of high variability between individuals, sequencing a substantial length of mitochondrial DNA provides a fingerprint that is highly specific to an individual.

Analysis of mitochondrial DNA is also important for evolutionary and epidemiological studies.

The reference sequence can be an entire mitochondrial genome or any fragment thereof. For forensic and epidemiological studies, the reference sequence is often all or part of the D-loop region in which variability between individuals is greatest (*e.g.*, from 16024-16401 and 29-408). For detection of mutations, analysis of the entire genome is useful as a reference sequence, but shorter segments including the sites of known mutations, and about 1-20 flanking bases are also useful. Some chips have probes tiling paired reference sequences, representing wildtype and mutant versions of a sequence. Tiling a second reference sequence is particularly useful for detecting an insertion mutation occurring in 30-50% of ocular myopathy and Pearson syndrome patients, which consists of direct repeats of the sequence ACCTCCCTCACCA. Some chips include reference sequences from more than one mitochondrial genome.

Mitochondrial reference sequences can be tiled using any of the strategies noted above. The block tiling strategy is particularly useful for analyzing short reference sequences or known mutations. Either the block strategy or the basic strategy is suitable for analyzing long reference sequences. In many of the tiling strategies, it is possible to use fewer probes compared with the number used in other chips without significant loss of sequence information. As noted above, most point mutations in mitochondrial DNA are transitions, so for each wildtype nucleotide in a reference sequence, one of the three possible nucleotide substitutions is much more likely than the other two. Accordingly, in the basic tiling strategy, for example, a reference sequence can be tiled using only two probe sets. One probe sets comprises a plurality of probes, each probe having a segment exactly complementary to the reference sequence. The second probe set comprises a corresponding probe for each probe in the first set. However, a probe from the second probe set differs from the corresponding probe from the first probe set in an interrogation position, in which the probe from the second

probe set includes the transition of the nucleotide present in that position in the probe from the first probe set.

Target mitochondrial DNA can be amplified, labelled and fragmented prior to hybridization using the same procedures as described for other chips. Use of at least two labelled nucleotides is desirable to achieve uniform labelling. Some exemplary primers are described below and other primers can be designed from the known sequence of mitochondrial DNA. Because mitochondrial DNA is present in multiple copies per cell, it can also be hybridized directly to a chip without prior amplification.

Exemplary Chips

The invention provides a DNA chip for analyzing sequences contained in a 1.3 kb fragment of human mitochondrial DNA from the "D-loop" region, the most polymorphic region of human mitochondrial DNA. One such chip comprises a set of 269 overlapping oligonucleotide probes of varying length in the range of 9-14 nucleotides with varying overlaps arranged in ~600 x 600 micron features or synthesis sites in an array 1 cm x 1 cm in size. The probes on the chip are shown in columnar form below. An illustrative mitochondrial DNA chip of the invention comprises the following probes (X, Y coordinates are shown, followed by the sequence; "DL3" represents the 3'-end of the probe, which is covalently attached to the chip surface.)

0	0	DL3AGTGGGGTATTT	1	1	DL3GGTTGGTTTGGG
1	0	DL3GGGTATTTAGTT	2	1	DL3TGGGGTTTCTAG
2	0	DL3TTAGTTTATCCAA	3	1	DL3GTTTCTAGTGGG
3	0	DL3ATCCAAACCAGG	4	1	DL3AGTGGGGGGTGT
4	0	DL3ACCAGGATCGGA	5	1	DL3GGGGTGTCAAAT
5	0	DL3CGTGTGTGTGTGG	6	1	DL3GTCAAATACATCG
6	0	DL3CGTGTGTGTGTGGC	7	1	DL3ACATCGAATGGAG
7	0	DL3TCGTGTGTGTGTGG	8	1	DL3CGAATGGAGGAG
8	0	DL3GTAGGATGGGTC	9	1	DL3GAGGAGTTTCGT
9	0	DL3AGGATGGGTCGT	10	1	DL3TTTCGTTATGTGA
10	0	DL3GATGGGTCGTGT	11	1	DL3ATGTGACTTTTAC
11	0	DL3TGGCGACGATTG	12	1	DL3GACTTTTACAAAT
12	0	DL3GCGACGATTGGG	13	1	DL3AAATCTGCCCGA
13	0	DL3TGGGGGGGA	14	1	DL3AATCTGCCCGAG
14	0	DL3GAGGGGGCG	15	1	DL3CCCGAGTGTAGT
15	0	DL3GGAGGGGGCGA	16	1	DL3AGTGTAGTGGGG
16	0	DL3GAGGGGGCGA	0	2	DL3GGGAGGGTGAG
0	1	DL3GGCTTGGTTGG	1	2	DL3GGTGAGGGTATG

2	2	DL3GGTATGATGATTAG	
3	2	DL3GATTAGAGTAAGT	
4	2	DL3TTAGAGTAAGTTA	
5	2	DL3AAGTTATGTTGGG	
6	2	DL3GTTGGGGGCG	
7	2	DL3GGGGCGGGTA	
8	2	DL3GCGGGTAGGAT	
9	2	DL3GGTAGGATGGGT	
10	2	DL3GGATGGGTCGTG	
11	2	DL3GGTCGTGTGTGT	
12	2	DL3GTGTGTGTGGCG	
13	2	DL3TGTGGCGACGAT	
14	2	DL3GACGATTGGGGT	
15	2	DL3ATTGGGGTATGG	
16	2	DL3GTATGGGGCTTG	
0	3	DL3GGATTGTGGTCG	
1	3	DL3TGGTCGGATTGG	
2	3	DL3GGATTGGTCTAAA	
3	3	DL3TCTAAAGTTTAAA	
4	3	DL3GTTTAAAATAGAA	
5	3	DL3ATAGAAAAACCG	
6	3	DL3AGAAAAACCGC	
7	3	DL3AACCGCCATAC	
8	3	DL3CCATACGTGAAAA	
9	3	DL3ACGTGAAAATTGT	
10	3	DL3AATTGTCAGTGGG	
11	3	DL3TGTCAGTGGGGG	
12	3	DL3TGGGGGGTTGA	
13	3	DL3GGGTTGATTGTGT	
14	3	DL3TTGTGTAATAAAA	
15	3	DL3AATAAAAGGGGA	
16	3	DL3TAAAAGGGGAGG	
0	4	DL3GTTTTTTAAAGG	
1	4	DL3TTTTAAAGGTGG	
2	4	DL3AGGTGGTTTGG	
3	4	DL3TTGGGGGGGAG	
4	4	DL3GGAGGGGGCG	
5	4	DL3GGGGCGAAGAC	
6	4	DL3GAAGACCGGATG	
7	4	DL3CCGGATGTCGTG	
8	4	DL3GTCGTGAATTTGT	
9	4	DL3CGTGAATTTGTGT	
10	4	DL3TTGTGTAGAGACG	
11	4	DL3TAGAGACGGTTT	
12	4	DL3ACGGTTTGGGG	
13	4	DL3TGGGGTTTTTGT	
14	4	DL3GGGTTTTTGTTT	
15	4	DL3TTGTTTCTTGGG	
16	4	DL3TCTTGGGATTGTG	
0	5	DL3TGTATGAATGATTT	
1	5	DL3TGATTTCACACAA	
2	5	DL3ACACAATTAATTAA	
3	5	DL3AATTAATTACGAA	
4	5	DL3TACGAACATCCTG	
5	5	DL3ACGAACATCCTGT	
6	5	DL3TCCTGTATTATTA	
7	5	DL3GTATTATTATTGTT	
8	5	DL3ATTGTTAAACTTA	
9	5	DL3AAACTTACAGACG	
10	5	DL3ACAGACGTGTCG	
11	5	DL3GTGTCGGTGAAA	
12	5	DL3GTGAAAGGTGTGT	
13	5	DL3GGTGTGTCTGTAG	
14	5	DL3TGTGTCTGTAGTA	
15	5	DL3GTAGTATTGTTTT	
16	5	DL3AGTATTGTTTTTT	
0	6	DL3CCTCGTGGGATA	
1	6	DL3TGGGATACAGCG	
2	6	DL3GATACAGCGTCAT	
3	6	DL3GCGTCATAGACAG	
4	6	DL3AGACAGAAACTAA	
5	6	DL3CAGAAACTAAGGA	
6	6	DL3TAAGGACGGAGT	
7	6	DL3GACGGAGTAGGA	
8	6	DL3GTAGGATAATAAA	
9	6	DL3TAATAAATAGCG	
10	6	DL3ATAGCGTAGGAT	
11	6	DL3TAGCGTAGGATG	
12	6	DL3AGGATGCAAGTT	
13	6	DL3ATGCAAGTTATAA	
14	6	DL3GTTATAATGTCCG	
15	6	DL3ATGTCCGCTTGT	
16	6	DL3TCCGCTTGTATG	
0	7	DL3GTGAGTGCCCTC	
1	7	DL3TGCCCTCGAGAG	
2	7	DL3CCTCGAGAGGTA	
3	7	DL3AGAGGTACGTAA	
4	7	DL3ACGTAAACCATA	
5	7	DL3ACCATAAAAGCAG	
6	7	DL3AAAGCAGACCC	
7	7	DL3AGACCCCCCAT	
8	7	DL3CCCCCATACGT	
9	7	DL3CATACGTGCGCT	
10	7	DL3GTGCGCTATCAG	
11	7	DL3GCGCTATCAGTA	
12	7	DL3TCAGTAACGCTC	
13	7	DL3GTAACGCTCTGC	
14	7	DL3CTCTGCGACCTC	
15	7	DL3GACCTCGGCCT	
16	7	DL3TCGGCCTCGTG	
0	8	DL3GATGAAGTCCCAG	
1	8	DL3AGTCCCAGTATTT	
2	8	DL3GTATTTCGGATTT	
3	8	DL3TCGGATTTATCG	
4	8	DL3GATTTATCGGGT	
5	8	DL3ATCGGGTGTGCA	
6	8	DL3TGTGCAAGGGGA	
7	8	DL3CAAGGGGAATTT	
8	8	DL3GAATTTATTCTGTA	
9	8	DL3TCTGTAGTGCTAC	
10	8	DL3GTAGTGCTACCT	
11	8	DL3GCTACCTAGTAG	
12	8	DL3CTAGTAGTCCAGA	
13	8	DL3TCCAGATAGTGGG	

14	8	DL3AGATAGTGGGATA	
15	8	DL3GGGATAATTGGT	
16	8	DL3TAATTGGTGAGTG	
0	9	DL3TATAGGGCGTGT	
1	9	DL3GGCGTGTTCTCA	
2	9	DL3GTGTTCTCACGAT	
3	9	DL3TCACGATGAGAGG	
4	9	DL3ATGAGAGGAGCG	
5	9	DL3AGGAGCGAGGC	
6	9	DL3CGAGGCCCGG	
7	9	DL3GCCCGGGTATT	
8	9	DL3CGGGTATTGTGA	
9	9	DL3GTGAACCCCCAT	
10	9	DL3CCCCATCGATTT	
11	9	DL3ATCGATTTCACTT	
12	9	DL3TTTCACTTGACAT	
13	9	DL3TTGACATAGAGCT	
14	9	DL3TAGAGCTGTAGAC	
15	9	DL3GTAGACCAAGGA	
16	9	DL3ACCAAGGATGAAG	
0	10	DL3CGTGTAATGTCAG	
1	10	DL3TGTCAGTTTAGGG	
2	10	DL3TCAGTTTAGGGA	
3	10	DL3TAGGGAAGAGCA	
4	10	DL3AAGAGCAGGGGT	
5	10	DL3CAGGGGTACCTA	
6	10	DL3GGTACCTACTGG	
7	10	DL3TACTGGGGGGA	
8	10	DL3GGGGGAGTCTAT	
9	10	DL3AGTCTATCCCCA	
10	10	DL3ATCCCCAGGGA	
11	10	DL3CAGGGAACTGGT	
12	10	DL3ACTGGTGGTAGG	
13	10	DL3CTGGTGGTAGGA	
14	10	DL3GTAGGAGGCACA	
15	10	DL3GGCACATTTAGT	
16	10	DL3TTTAGTTATAGGG	
0	11	DL3AGGTTTACGGTG	
1	11	DL3TACGGTGGGGA	
2	11	DL3GTGGGGAGTGG	
3	11	DL3GGGAGTGGGTGA	
4	11	DL3GGGTGATCCTATG	
5	11	DL3CCTATGGTTGTTT	
6	11	DL3GGTTGTTTGGATG	
7	11	DL3GTTTGGATGGGT	
8	11	DL3ATGGGTGGGAAT	
9	11	DL3GGGAATTGTCATG	
10	11	DL3GTCATGTATCATGT	
11	11	DL3TCATGTATTTCGG	
12	11	DL3TATTTCGGTAAA	
13	11	DL3TTCGGTAAATGG	
14	11	DL3GTAAATGGCATGT	
15	11	DL3GCATGTAATCGTG	
16	11	DL3GTAATCGTGTAAT	
5	12	DL3GGGAGGGGTAC	
6	12	DL3GGGTACGAATGT	
7	12	DL3ACGAATGTTCGTT	
8	12	DL3TGTTCGTTCATGT	
9	12	DL3CGTTCATGTCGTT	
10	12	DL3GTCGTTAGTTGG	
11	12	DL3TAGTTGGGAGTT	
12	12	DL3GGAGTTGATAGTG	
13	12	DL3ATAGTGTGTAGTT	
14	12	DL3GTGTAGTTGACGT	
15	12	DL3TGACGTTGAGGT	
16	12	DL3CGTTGAGGTTTA	
5	13	DL3TATAACATGCCAT	
6	13	DL3AACATGCCATGGT	
7	13	DL3CCATGGTATTTAT	
8	13	DL3ATTTATGAACTGG	
9	13	DL3AACTGGTGGACAT	
10	13	DL3TGGACATCATGTA	
11	13	DL3CATGTATTTTTGG	
12	13	DL3TTTTGGGTTAGG	
13	13	DL3GGGTTAGGATGT	
14	13	DL3GGATGTAGTTTTG	
15	13	DL3TGTAGTTTTGGG	
16	13	DL3TTTGGGGGAGG	
5	14	DL3GGGTTCATAACTG	
6	14	DL3ATAACTGAGTGGG	
7	14	DL3AACTGAGTGGGT	
8	14	DL3GTGGGTAGTTGT	
9	14	DL3GTAGTTGTTGGC	
10	14	DL3GTTGGCGATACA	
11	14	DL3CGATACATAAAAG	
12	14	DL3TAAAAGCATGTAA	
13	14	DL3GCATGTAATGACG	
14	14	DL3ATGACGGTCGGT	
15	14	DL3GTCGGTGGTACT	
16	14	DL3GGTACTTATAACA	
5	15	DL3TCGATTCTAAGAT	
6	15	DL3TAAGATTAAATTT	
7	15	DL3AAATTTGAATAAG	
8	15	DL3AATAAGAGACAAG	
9	15	DL3AAGAGACAAGAAA	
10	15	DL3AAGAAAGTACCC	
11	15	DL3AAAGTACCCCTT	
12	15	DL3CCCCTTCGTCTA	
13	15	DL3CTTCGTCTAAAC	
14	15	DL3CTAAACCCATGG	
15	15	DL3AACCCATGGTGG	
16	15	DL3TGGTGGGTTCAT	
5	16	DL3TTGGAAAAAGGT	
6	16	DL3AAAAGGTTCCTG	
7	16	DL3GGTTCCTGTTTA	
8	16	DL3CCTGTTTAGTCTC	
9	16	DL3TTAGTCTCTTTTT	
10	16	DL3CTTTTTCAGAAAT	
11	16	DL3AGAAATTGAGGTG	
12	16	DL3AAATTGAGGTGGT	
13	16	DL3GGTGGTAATCGT	
14	16	DL3TAATCGTGGGTT	
15	16	DL3GTGGGTTTCGAT	
16	16	DL3GGTTTCGATTCT	

determined from 27 bright features. After scanning, the chip was stripped and rehybridized; all six samples were hybridized to the same chip. Figure 36 shows the image observed from the mt4 sample on the DNA chip. Figure 37 shows the image observed from the mt5 sample on the DNA chip. Figure 38 shows the predicted difference image between the mt4 and mt5 samples on the DNA chip based on mismatches between the two samples and the reference sequence (*see* Anderson et al., *supra*). Figure 39 shows the actual difference image observed.

The results show that, in almost all cases, mismatched probe/target hybrids resulted in lower fluorescence intensity than perfectly matched hybrids. Nonetheless, some probes detected mutations (or specific sequences) better than others, and in several cases, the differences were within noise levels. Improvements can be realized by increasing the amount of overlap between probes and hence overall probe density and, for duplex DNA targets, using a second set of probes, either on the same or a separate chip, corresponding to the second strand of the target. Figure 40, in sheets 1 and 2, shows a plot of normalized intensities across rows 10 and 11 of the array and a tabulation of the mutations detected.

Figure 41 shows the discrimination between wild-type and mutant hybrids obtained with this chip. The median of the six normalized hybridization scores for each probe was taken. The graph plots the ratio of the median score to the normalized hybridization score versus mean counts. On this graph, a ratio of 1.6 and mean counts above 50 yield no false positives, and while it is clear that detection of some mutants can be improved, excellent discrimination is achieved, considering the small size of the array. Figure 42 illustrates how the identity of the base mismatch may influence the ability to discriminate mutant and wild-type sequences more than the position of the mismatch within an oligonucleotide probe. The mismatch position is expressed as % of probe length from the 3'-end. The base change is indicated on the graph. These results show that the DNA chip increases the capacity of the standard reverse dot blot format by orders of magnitude, extending the power of that approach

many fold and that the methods of the invention are more efficient and easier to automate than gel-based methods of nucleic acid sequence and mutation analysis.

To illustrate further these advantages, a second chip was prepared for analyzing a longer segment from human mitochondrial DNA (mtDNA). The chip "tiles" through 648 nucleotides of a reference sequence comprising human H strand mtDNA from positions 16280 to 356, and allows analysis of each nucleotide in the reference sequence. The probes in the array are 15 nucleotides in length, and each position in the target sequence is represented by a set of 4 probes (A, C, G, T substitutions), which differed from one another at position 7 from the 3'-end. The array consists of 13 blocks of 4 x 50 probes: each block scans through 50 nucleotides of contiguous mtDNA sequence. The blocks are separated by blank rows. The 4 corner columns contain control probes; there are a total of 2600 probes in a 1.28 cm x 1.28 cm square area (feature), and each area is 256 x 197 microns.

Target RNA was prepared as above. The RNA was fragmented and hybridized to the oligonucleotide array in a solution composed of 6X SSPE, 0.1% Triton X-100 for 60 minutes at 18°C. Unhybridized material was washed away with buffer, and the chip was scanned at 25 micron pixel resolution.

Figure 43 provides a 5' to 3' sequence listing of one target corresponding to the probes on the chip. X is a control probe. Positions that differ in the target (*i.e.*, are mismatched with the probe at the designated site) are in bold. Figure 44 shows the fluorescence image produced by scanning the chip when hybridized to this sample. About 95% of the sequence could be read correctly from only one strand of the original duplex target nucleic acid. Although some probes did not provide excellent discrimination and some probes did not appear to hybridize to the target efficiently, excellent results were achieved. The target sequence differed from the probe set at six positions: 4 transitions and 2 insertions. All 4 transitions were detected, and specific probes could readily be incorporated into the array to detect insertions or deletions. Figure 45 illustrates the detection of 4

transitions in the target sequence relative to the wild-type probes on the chip.

A further chip was constructed comprising probes tiling across the entire D-loop region (1.3 kb) of mt DNA sequences from two humans. The probes were tiled in rows of four using the basic tiling strategy. The probes were overlapping 15 mers having an interrogation position 7 nucleotides from the 3' end. The complete group of probes tiled on the reference sequence from the first individual, designated mt1, occupied the upper half of the chip. The lower half of the chip contained a similar arrangement based on a second clone, mt2. The probes were synthesized in a 1.28 x 1.28 cm area, which contained a matrix of 115 x 120 cells. The chip contained a total of 10,488 mtDNA probes.

Six samples of target DNA was extracted form hair roots from six individuals. The 1.3 kb region spanning positions 15935 to 667 of human mtDNA was PCR amplified, cloned in bacteriophage M13 and sequenced by conventional methods. The 1.3 kb region was reamplified from the phage clone using primers, L15935-T3, 5'CTCGGAATTAACCCTCACTAAAGGAAACCTTTTTCCAAGGA and H667-T7, 5'TAATACGACTCACTATAGGGAGAGGCTAGGACCAAACCTATT tagged with T3 and T7 RNA polymerase promoter sequences. Labelled RNA was generated by *in vitro* transcription using T3 RNA polymerase and fluoresceinated nucleotides, fragmented, and hybridized to the mtDNA control region resequencing chip at room temperature for 60 min, in 6xSSPE + 0.05% triton X-100. Six washes were carried out at room temperature,using 6xSSPE + 0.005% triton X-100, and the chip was read. Signal intensities varied considerably over the chip, but the large dynamic range of the detection system allowed accurate quantitation of intensities over several orders of magnitude. Even relatively low signal intensities yielded accurate results.

Five different clones (mt1-5) were hybridized, each to a separate chip. The reference sequence was also hybridized for comparative purposes. Mean counts per probe cell were determined, and used by automated basecalling software to read the sequence. The accuracy of sequence read from the chip is

summarized as follows. Combining the data from the five targets analyzed, the chip read a total of 6310 nucleotides. Of these nucleotides in the target sequences, 55 were different from the reference sequence (as judged by conventional sequencing) (41 of these 55 nucleotides were both detected and read correctly from the chip). 6 of 55 nucleotides were detected as being ambiguous but their identity could not be read. 2 of 55 nucleotides were detected as mutations, but their identity was miscalled. 6 of 55 nucleotides were incorrectly called as wildtype. Of the 6255 nucleotides in the target sequence that were identical to the reference sequence, only 36 (0.57%) were miscalled or scored as ambiguous.

A further chip was constructed comprising probes tiling across a reference sequence comprising an entire mitochondrial genome. In this chip, a block tiling strategy was used. Each block was designed to analyze seven nucleotides from a target sequence. Each block consisted of four probe sets, the probe sets each having seven probes. A block was laid down on the chip in seven columns of four probes. The upper probe was the same in each column, this being a probe exactly complementary to a subsequence of the reference sequence. The three other probes in each column were identical to the upper probe except in an interrogation position, which was occupied by a different base in each of the four probes in the column. The interrogation position shifted by one position between successive columns. Thus, except for the seven interrogation positions, one in each of the columns of probes, all probes occupying a block were identical. The array comprised many such blocks, each tiled to successive subsequences of the mitochondrial DNA reference sequence. In all, the chip tiled 15,569 nucleotides of reference sequence with double tiling at 42 positions. 66,276 probes occupied an array of 304 x 315 cells, each cell having an area of 42 x 41 microns.

The chip was hybridized to the same target sequences as described for the D-loop region, except that hybridization was at 15°C for 2 hr. The chip was scanned at 5 micron resolution to give an image with approximately 64 pixels per cell. For

blocks of probes tiling across the D-loop region, a sequence-specific hybridization pattern was obtained. For other blocks, only background hybridization was observed.

These results illustrate that longer sequences can be read using the DNA chips and methods of the invention, as compared to conventional sequencing methods, where reading length is limited by the resolution of gel electrophoresis. Hybridization and signal detection require less than an hour and can be readily shortened by appropriate choice of buffers, temperatures, probes, and reagents.

III. MODES OF PRACTICING THE INVENTION

A. VLSIPS™ Technology

As noted above, the VLSIPS™ technology is described in a number of patent publications and is preferred for making the oligonucleotide arrays of the invention. A brief description of how this technology can be used to make and screen DNA chips is provided in this Example and the accompanying Figures. In the VLSIPS™ method, light is shone through a mask to activate functional (for oligonucleotides, typically an -OH) groups protected with a photoremovable protecting group on a surface of a solid support. After light activation, a nucleoside building block, itself protected with a photoremovable protecting group (at the 5'-OH), is coupled to the activated areas of the support. The process can be repeated, using different masks or mask orientations and building blocks, to prepare very dense arrays of many different oligonucleotide probes. The process is illustrated in Figure 46; Figure 47 illustrates how the process can be used to prepare "nucleoside combinatorials" or oligonucleotides synthesized by coupling all four nucleosides to form dimers, trimers and so forth.

New methods for the combinatorial chemical synthesis of peptide, polycarbamate, and oligonucleotide arrays have recently been reported (*see* Fodor et al., 1991, *Science* 251: 767-773; Cho et al., 1993, *Science* 261: 1303-1305; and Southern et al., 1992, *Genomics* 13: 1008-10017, each of which is incorporated herein by reference). These arrays, or

biological chips (*see* Fodor et al., 1993, *Nature* 364: 555-556, incorporated herein by reference), harbor specific chemical compounds at precise locations in a high-density, information rich format, and are a powerful tool for the study of biological recognition processes. A particularly exciting application of the array technology is in the field of DNA sequence analysis. The hybridization pattern of a DNA target to an array of shorter oligonucleotide probes is used to gain primary structure information of the DNA target. This format has important applications in sequencing by hybridization, DNA diagnostics and in elucidating the thermodynamic parameters affecting nucleic acid recognition.

Conventional DNA sequencing technology is a laborious procedure requiring electrophoretic size separation of labeled DNA fragments. An alternative approach, termed Sequencing By Hybridization (SBH), has been proposed (Lysov et al., 1988, *Dokl. Akad. Nauk SSSR* 303:1508-1511; Bains et al., 1988, *J. Theor. Biol.* 135:303-307; and Drmanac et al., 1989, *Genomics* 4:114-128, incorporated herein by reference). This method uses a set of short oligonucleotide probes of defined sequence to search for complementary sequences on a longer target strand of DNA. The hybridization pattern is used to reconstruct the target DNA sequence. It is envisioned that hybridization analysis of large numbers of probes can be used to sequence long stretches of DNA. In immediate applications of this hybridization methodology, a small number of probes can be used to interrogate local DNA sequence.

The strategy of SBH can be illustrated by the following example. A 12-mer target DNA sequence, AGCCTAGCTGAA, is mixed with a complete set of octanucleotide probes. If only perfect complementarity is considered, five of the 65,536 octamer probes -TCGGATCG, CGGATCGA, GGATCGAC, GATCGACT, and ATCGACTT will hybridize to the target. Alignment of the overlapping sequences from the hybridizing probes reconstructs the complement of the original 12-mer target:

```
TCGGATCG
 CGGATCGA
  GGATCGAC
   GATCGACT
```

ATCGACTT
TCGGATCGACTT

Hybridization methodology can be carried out by attaching target DNA to a surface. The target is interrogated with a set of oligonucleotide probes, one at a time (*see* Strezoska et al., 1991, *Proc. Natl. Acad. Sci. USA* 88:10089-10093, and Drmanac et al., 1993, *Science* 260:1649-1652, each of which is incorporated herein by reference). This approach can be implemented with well established methods of immobilization and hybridization detection, but involves a large number of manipulations. For example, to probe a sequence utilizing a full set of octanucleotides, tens of thousands of hybridization reactions must be performed. Alternatively, SBH can be carried out by attaching probes to a surface in an array format where the identity of the probes at each site is known. The target DNA is then added to the array of probes. The hybridization pattern determined in a single experiment directly reveals the identity of all complementary probes.

As noted above, a preferred method of oligonucleotide probe array synthesis involves the use of light to direct the synthesis of oligonucleotide probes in high-density, miniaturized arrays. Photolabile 5'-protected N-acyl-deoxynucleoside phosphoramidites, surface linker chemistry, and versatile combinatorial synthesis strategies have been developed for this technology. Matrices of spatially-defined oligonucleotide probes have been generated, and the ability to use these arrays to identify complementary sequences has been demonstrated by hybridizing fluorescent labeled oligonucleotides to the DNA chips produced by the methods. The hybridization pattern demonstrates a high degree of base specificity and reveals the sequence of oligonucleotide targets.

The basic strategy for light-directed oligonucleotide synthesis (1) is outlined in Fig. 46. The surface of a solid support modified with photolabile protecting groups (X) is illuminated through a photolithographic mask, yielding reactive hydroxyl groups in the illuminated regions. A 3'-O-phosphoramidite activated deoxynucleoside (protected at

the 5'-hydroxyl with a photolabile group) is then presented to the surface and coupling occurs at sites that were exposed to light. Following capping, and oxidation, the substrate is rinsed and the surface illuminated through a second mask, to expose additional hydroxyl groups for coupling. A second 5'-protected, 3'-O-phosphoramidite activated deoxynucleoside is presented to the surface. The selective photodeprotection and coupling cycles are repeated until the desired set of products is obtained.

Light directed chemical synthesis lends itself to highly efficient synthesis strategies which will generate a maximum number of compounds in a minimum number of chemical steps. For example, the complete set of 4^n polynucleotides (length n), or any subset of this set can be produced in only 4 x n chemical steps. *See* Fig. 47. The patterns of illumination and the order of chemical reactants ultimately define the products and their locations. Because photolithography is used, the process can be miniaturized to generate high-density arrays of oligonucleotide probes. For an example of the nomenclature useful for describing such arrays, an array containing all possible octanucleotides of dA and dT is written as $(A+T)^8$. Expansion of this polynomial reveals the identity of all 256 octanucleotide probes from AAAAAAAA to TTTTTTTT. A DNA array composed of complete sets of dinucleotides is referred to as having a complexity of 2. The array given by (A+T+C+G)8 is the full 65,536 octanucleotide array of complexity four. Computer-aided methods of laying down predesigned arrays of probes using VLSIPS™ technology are described in commonly-assigned co-pending application USSN 08/249,188, filed May 24, 1994 (incorporated by reference in its entirety for all purposes).

To carry out hybridization of DNA targets to the probe arrays, the arrays are mounted in a thermostatically controlled hybridization chamber. Fluorescein labeled DNA targets are injected into the chamber and hybridization is allowed to proceed for 5 min to 24 hr. The surface of the matrix is scanned in an epifluorescence microscope (Zeiss Axioscop 20) equipped with photon counting electronics using

50 - 100 μW of 488 nm excitation from an Argon ion laser (Spectra Physics Model 2020). Measurements may be made with the target solution in contact with the probe matrix or after washing. Photon counts are stored and image files are presented after conversion to an eight bit image format. *See* Fig. 51.

When hybridizing a DNA target to an oligonucleotide array, N = Lt-(Lp-1) complementary hybrids are expected, where N is the number of hybrids, Lt is the length of the DNA target, and Lp is the length of the oligonucleotide probes on the array. For example, for an 11-mer target hybridized to an octanucleotide array, N = 4. Hybridizations with mismatches at positions that are 2 to 3 residues from either end of the probes will generate detectable signals. Modifying the above expression for N, one arrives at a relationship estimating the number of detectable hybridizations (Nd) for a DNA target of length Lt and an array of complexity C. Assuming an average of 5 positions giving signals above background:

$Nd = (1 + 5(C-1))[Lt-(Lp-1)]$.

Arrays of oligonucleotides can be efficiently generated by light-directed synthesis and can be used to determine the identity of DNA target sequences. Because combinatorial strategies are used, the number of compounds increases exponentially while the number of chemical coupling cycles increases only linearly. For example, synthesizing the complete set of 4^8 (65,536) octanucleotides will add only four hours to the synthesis for the 16 additional cycles. Furthermore, combinatorial synthesis strategies can be implemented to generate arrays of any desired composition. For example, because the entire set of dodecamers (4^{12}) can be produced in 48 photolysis and coupling cycles (b^n compounds requires b x n cycles), any subset of the dodecamers (including any subset of shorter oligonucleotides) can be constructed with the correct lithographic mask design in 48 or fewer chemical coupling steps. In addition, the number of compounds in an array is limited only by the density of synthesis sites and the overall array size. Recent experiments have demonstrated hybridization to probes

synthesized in 25 μm sites. At this resolution, the entire set of 65,536 octanucleotides can be placed in an array measuring 0.64 cm square, and the set of 1,048,576 dodecanucleotides requires only a 2.56 cm array.

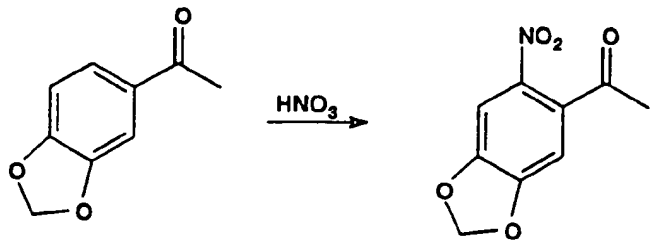
Genome sequencing projects will ultimately be limited by DNA sequencing technologies. Current sequencing methodologies are highly reliant on complex procedures and require substantial manual effort. Sequencing by hybridization has the potential for transforming many of the manual efforts into more efficient and automated formats. Light-directed synthesis is an efficient means for large scale production of miniaturized arrays for SBH. The oligonucleotide arrays are not limited to primary sequencing applications. Because single base changes cause multiple changes in the hybridization pattern, the oligonucleotide arrays provide a powerful means to check the accuracy of previously elucidated DNA sequence, or to scan for changes within a sequence. In the case of octanucleotides, a single base change in the target DNA results in the loss of eight complements, and generates eight new complements. Matching of hybridization patterns may be useful in resolving sequencing ambiguities from standard gel techniques, or for rapidly detecting DNA mutational events. The potentially very high information content of light-directed oligonucleotide arrays will change genetic diagnostic testing. Sequence comparisons of hundreds to thousands of different genes will be assayed simultaneously instead of the current one, or few at a time format. Custom arrays can also be constructed to contain genetic markers for the rapid identification of a wide variety of pathogenic organisms.

Oligonucleotide arrays can also be applied to study the sequence specificity of RNA or protein-DNA interactions. Experiments can be designed to elucidate specificity rules of non Watson-Crick oligonucleotide structures or to investigate the use of novel synthetic nucleoside analogs for antisense or triple helix applications. Suitably protected RNA monomers may be employed for RNA synthesis. The oligonucleotide arrays should find broad application deducing the thermodynamic and

kinetic rules governing formation and stability of oligonucleotide complexes.

Other than the use of photoremovable protecting groups, the nucleoside coupling chemistry is very similar to that used routinely today for oligonucleotide synthesis. Fig. 48 shows the deprotection, coupling, and oxidation steps of a solid phase DNA synthesis method. Fig. 49 shows an illustrative synthesis route for the nucleoside building blocks used in the method. Fig. 50 shows a preferred photoremovable protecting group, MeNPOC, and how to prepare the group in active form. The procedures described below show how to prepare these reagents. The nucleoside building blocks are 5'-MeNPOC-THYMIDINE-3'-OCEP; 5'-MeNPOC-N^4-t-BUTYL PHENOXYACETYL-DEOXYCYTIDINE-3'-OCEP; 5'-MeNPOC-N^4-t-BUTYL PHENOXYACETYL-DEOXYGUANOSINE-3'-OCEP; and 5'-MeNPOC-N^4-t-BUTYL PHENOXYACETYL-DEOXYADENOSINE-3'-OCEP..

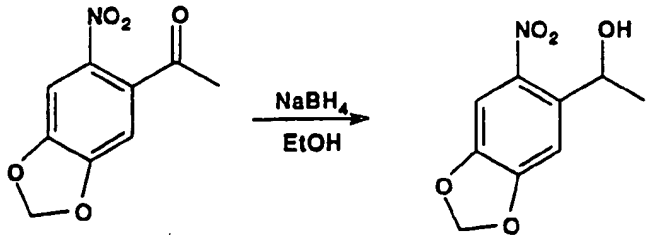
1. Preparation of 4,5-methylenedioxy-2-nitroacetophenone



A solution of 50 g (0.305 mole) 3,4-methylenedioxy-acetophenone (Aldrich) in 200 mL glacial acetic acid was added dropwise over 30 minutes to 700 mL of cold (2-4°C) 70% HNO_3 with stirring (NOTE: the reaction will overheat without external cooling from an ice bath, which can be dangerous and lead to side products). At temperatures below 0°C, however, the reaction can be sluggish. A temperature of 3-5°C seems to be optimal). The mixture was left stirring for another 60 minutes at 3-5°C, and then allowed to approach ambient temperature. Analysis by TLC (25% EtOAc in hexane) indicated complete conversion of the starting material within 1-2 hr. When the reaction was complete, the mixture was poured into ~3 liters of crushed ice, and the resulting yellow solid was

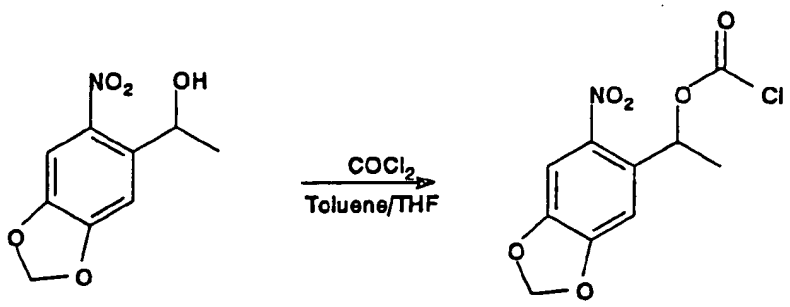
filtered off, washed with water and then suction-dried. Yield ~53 g (84%), used without further purification.

2. Preparation of 1-(4,5-Methylenedioxy-2-nitrophenyl) ethanol



Sodium borohydride (10g; 0.27 mol) was added slowly to a cold, stirring suspension of 53g (0.25 mol) of 4,5-methylenedioxy-2-nitroacetophenone in 400 mL methanol. The temperature was kept below 10°C by slow addition of the $NaBH_4$ and external cooling with an ice bath. Stirring was continued at ambient temperature for another two hours, at which time TLC (CH_2Cl_2) indicated complete conversion of the ketone. The mixture was poured into one liter of ice-water and the resulting suspension was neutralized with ammonium chloride and then extracted three times with 400 mL CH_2Cl_2 or EtOAc (the product can be collected by filtration and washed at this point, but it is somewhat soluble in water and this results in a yield of only ~60%). The combined organic extracts were washed with brine, then dried with $MgSO_4$ and evaporated. The crude product was purified from the main byproduct by dissolving it in a minimum volume of CH_2Cl_2 or THF (~175 ml) and then precipitating it by slowly adding hexane (1000 ml) while stirring (yield 51g; 80% overall). It can also be recrystallized (e.g., toluene-hexane), but this reduces the yield.

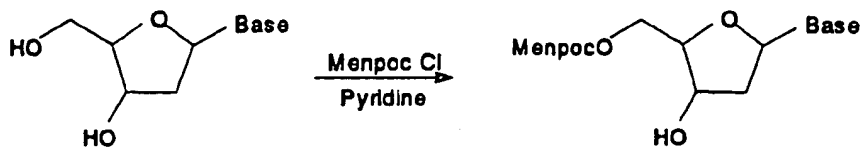
3. Preparation of 1-(4,5- methylenedioxy-2-nitrophenyl) ethyl chloroformate (MeNPOC-Cl)



Phosgene (500 mL of 20% w/v in toluene from Fluka: 965 mmole; 4 eq.) was added slowly to a cold, stirring solution of 50g (237 mmole; 1 eq.) of 1-(4,5-methylenedioxy-2-nitrophenyl) ethanol in 400 mL dry THF. The solution was stirred overnight at ambient temperature at which point TLC (20% Et_2O/hexane) indicated >95% conversion. The mixture was evaporated (an oil-less pump with downstream aqueous NaOH trap is recommended to remove the excess phosgene) to afford a viscous brown oil. Purification was effected by flash chromatography on a short (9 x 13 cm) column of silica gel eluted with 20% Et_2O/hexane. Typically 55g (85%) of the solid yellow MeNPOC-Cl is obtained by this procedure. The crude material has also been recrystallized in 2-3 crops from 1:1 ether/hexane. On this scale, ~100ml is used for the first crop, with a few percent THF added to aid dissolution, and then cooling overnight at -20°C (this procedure has not been optimized). The product should be stored desiccated at -20°C.

4. Synthesis of 5'- Menpoc-2'-deoxynucleoside-3'-(N,N-diisopropyl 2-cyanoethyl phosphoramidites

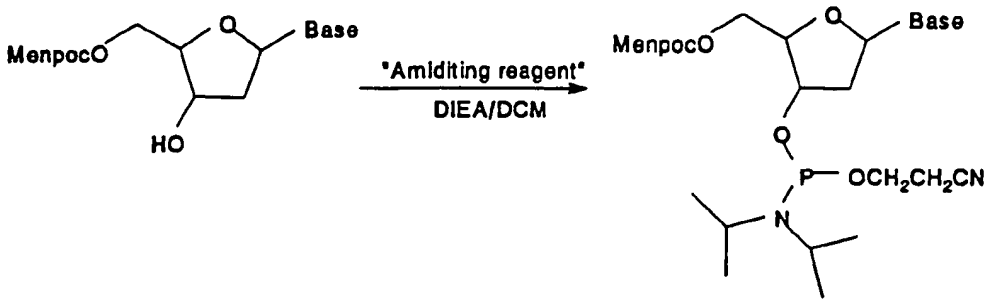
(a.) 5'-MeNPOC-Nucleosides



Base= THYMIDINE (T); N-4-ISOBUTYRYL 2'-DEOXYCYTIDINE (ibu-dC); N-2-PHENOXYACETYL 2'DEOXYGUANOSINE (PAC-dG); and N-6-PHENOXYACETYL 2'DEOXYADENOSINE (PAC-dA)

All four of the 5'-MeNPOC nucleosides were prepared from the base-protected 2'-deoxynucleosides by the following procedure. The protected 2'-deoxynucleoside (90 mmole) was dried by co-evaporating twice with 250 mL anhydrous pyridine. The nucleoside was then dissolved in 300 mL anhydrous pyridine (or 1:1 pyridine/DMF, for the dG^{PAC} nucleoside) under argon and cooled to ~2°C in an ice bath. A solution of 24.6g (90 mmole) MeNPOC-Cl in 100 mL dry THF was then added with stirring over 30 minutes. The ice bath was removed, and the solution allowed to stir overnight at room temperature (TLC: 5-10% MeOH in CH_2Cl_2, two diastereomers). After evaporating the solvents under vacuum, the crude material was taken up in 250 mL ethyl acetate and extracted with saturated aqueous $NaHCO_3$ and brine. The organic phase was then dried over Na_2SO_4, filtered and evaporated to obtain a yellow foam. The crude products were finally purified by flash chromatography (9 x 30 cm silica gel column eluted with a stepped gradient of 2% - 6% MeOH in CH_2Cl_2). Yields of the purified diastereomeric mixtures are in the range of 65-75%.

(b.) 5'- Menpoc-2'-deoxynucleoside-3'-(N,N-diisopropyl 2-cyanoethyl phosphoramidites)



The four deoxynucleosides were phosphitylated using either 2-cyanoethyl- N,N- diisopropyl chlorophosphoramidite, or 2-cyanoethyl- N,N,N',N'- tetraisopropylphosphorodiamidite. The following is a typical procedure. Add 16.6g (17.4 ml; 55 mmole) of 2- cyanoethyl- N,N,N',N'- tetraisopropylphosphoro-diamidite to a solution of 50 mmole 5'- MeNPOC-nucleoside and 4.3g (25 mmole) diisopropylammonium tetrazolide in 250 mL dry CH_2Cl_2 under argon at ambient temperature. Continue stirring for 4-16 hours (reaction monitored by TLC: 45:45:10 hexane/CH_2Cl_2/Et_3N). Wash the organic phase with saturated aqueous $NaHCO_3$ and brine, then dry over Na_2SO_4, and evaporate to dryness. Purify the crude amidite by flash chromatography (9 x 25 cm silica gel column eluted with hexane/CH_2Cl_2/TEA - 45:45:10 for A, C, T; or 0:90:10 for G). The yield of purified amidite is about 90%.

B. PREPARATION OF LABELED DNA/HYBRIDIZATION TO ARRAY

1. PCR

PCR amplification reactions are typically conducted in a mixture composed of, per reaction: 1 μl genomic DNA; 10 μl each primer (10 pmol/μl stocks); 10 μl 10 x PCR buffer (100 mM Tris.Cl pH8.5, 500 mM KCl, 15 mM $MgCl_2$); 10 μl 2 mM dNTPs (made from 100 mM dNTP stocks); 2.5 U Taq polymerase (Perkin Elmer AmpliTaq™, 5 U/μl); and H_2O to 100 μl. The cycling conditions are usually 40 cycles (94°C 45 sec, 55°C 30 sec, 72°C 60 sec) but may need to be varied considerably from

sample type to sample type. These conditions are for 0.2 mL thin wall tubes in a Perkin Elmer.9600 thermocycler. *See* Perkin Elmer 1992/93 catalogue for 9600 cycle time information. Target, primer length and sequence composition, among other factors, may also affect parameters.

For products in the 200 to 1000 bp size range, check 2 μl of the reaction on a 1.5% 0.5x TBE agarose gel using an appropriate size standard (phiX174 cut with *Hae*III is convenient). The PCR reaction should yield several picomoles of product. It is helpful to include a negative control (*i.e.*, 1 μl TE instead of genomic DNA) to check for possible contamination. To avoid contamination, keep PCR products from previous experiments away from later reactions, using filter tips as appropriate. Using a set of working solutions and storing master solutions separately is helpful, so long as one does not contaminate the master stock solutions.

For simple amplifications of short fragments from genomic DNA it is, in general, unnecessary to optimize Mg^{2+} concentrations. A good procedure is the following: make a master mix minus enzyme; dispense the genomic DNA samples to individual tubes or reaction wells; add enzyme to the master mix; and mix and dispense the master solution to each well, using a new filter tip each time.

2. PURIFICATION

Removal of unincorporated nucleotides and primers from PCR samples can be accomplished using the Promega Magic PCR Preps DNA purification kit. One can purify the whole sample, following the instructions supplied with the kit (proceed from section IIIB, 'Sample preparation for direct purification from PCR reactions'). After elution of the PCR product in 50 μl of TE or H_2O, one centrifuges the eluate for 20 sec at 12,000 rpm in a microfuge and carefully transfers 45 μl to a new microfuge tube, avoiding any visible pellet. Resin is sometimes carried over during the elution step. This transfer prevents accidental contamination of the linear amplification reaction with 'Magic PCR' resin. Other methods, *e.g.*, size exclusion chromatography, may also be used.

3. Linear amplification

In a 0.2 mL thin-wall PCR tube mix: 4 μl purified PCR product; 2 μl primer (10 pmol/μl); 4 μl 10 x PCR buffer; 4 μl dNTPs (2 mM dA, dC, dG, 0.1 mM dT); 4 μl 0.1 mM dUTP; 1 μl 1 mM fluorescein dUTP (Amersham RPN 2121); 1 U Taq polymerase (Perkin Elmer, 5 U/μl); and add H2O to 40 μl. Conduct 40 cycles (92°C 30 sec, 55°C 30 sec, 72°C 90 sec) of PCR. These conditions have been used to amplify a 300 nucleotide mitochondrial DNA fragment but are applicable to other fragments. Even in the absence of a visible product band on an agarose gel, there should still be enough product to give an easily detectable hybridization signal. If one is not treating the DNA with uracil DNA glycosylase (see Section 4), dUTP can be omitted from the reaction.

4. Fragmentation

Purify the linear amplification product using the Promega Magic PCR Preps DNA purification kit, as per Section 2 above. In a 0.2 mL thin-wall PCR tube mix: 40 μl purified labeled DNA; 4 μl 10 x PCR buffer; and 0.5 μl uracil DNA glycosylase (BRL 1U/μl). Incubate the mixture 15 min at 37°C, then 10 min at 97°C; store at -20°C until ready to use.

5. Hybridization, Scanning & Stripping

A blank scan of the slide in hybridization buffer only is helpful to check that the slide is ready for use. The buffer is removed from the flow cell and replaced with 1 mL of (fragmented) DNA in hybridization buffer and mixed well. The scan is performed in the presence of the labeled target. Fig. 51 illustrates an illustrative detection system for scanning a DNA chip. A series of scans at 30 min intervals using a hybridization temperature of 25°C yields a very clear signal, usually in at least 30 min to two hours, but it may be desirable to hybridize longer, *i.e.*, overnight. Using a laser power of 50 μW and 50 μm pixels, one should obtain maximum counts in the range of hundreds to low thousands/pixel for a new slide. When finished, the slide can be stripped using 50%

formamide. rinsing well in deionized H_2O, blowing dry, and storing at room temperature.

C. PREPARATION OF LABELED RNA/HYBRIDIZATION TO ARRAY

1. Tagged primers

The primers used to amplify the target nucleic acid should have promoter sequences if one desires to produce RNA from the amplified nucleic acid. Suitable promoter sequences are shown below and include:

(1) the T3 promoter sequence:

5'-CGGAATTAACCCTCACTAAAGG

5'-AATTAACCCTCACTAAAGGGAG;

(2) the T7 promoter sequence:

5' TAATACGACTCACTATAGGGAG;

and (3) the SP6 promoter sequence:

5' ATTTAGGTGACACTATAGAA.

The desired promoter sequence is added to the 5' end of the PCR primer. It is convenient to add a different promoter to each primer of a PCR primer pair so that either strand may be transcribed from a single PCR product.

Synthesize PCR primers so as to leave the DMT group on. DMT-on purification is unnecessary for PCR but appears to be important for transcription. Add 25 µl 0.5M NaOH to collection vial prior to collection of oligonucleotide to keep the DMT group on. Deprotect using standard chemistry -- 55°C overnight is convenient.

HPLC purification is accomplished by drying down the oligonucleotides, resuspending in 1 mL 0.1 M TEAA (dilute 2.0 M stock in deionized water, filter through 0.2 micron filter) and filter through 0.2 micron filter. Load 0.5 mL on reverse phase HPLC (column can be a Hamilton PRP-1 semi-prep, #79426). The gradient is 0 -> 50% CH_3CN over 25 min (program 0.2 µmol.prep.0-50, 25 min). Pool the desired fractions, dry down, resuspend in 200 µl 80% HAc. 30 min RT. Add 200 µl EtOH; dry down. Resuspend in 200 µl H_2O, plus 20 µl NaAc pH5.5, 600 µl EtOH. Leave 10 min on ice; centrifuge 12,000 rpm for 10 min in microfuge. Pour off supernatant. Rinse pellet with 1 mL

EtOH, dry, resuspend in 200 µl H2O. Dry, resuspend in 200 µl TE. Measure A260, prepare a 10 pmol/µl solution in TE (10 mM Tris.Cl pH 8.0, 0.1 mM EDTA). Following HPLC purification of a 42 mer, a yield in the vicinity of 15 nmol from a 0.2 µmol scale synthesis is typical.

2. Genomic DNA Preparation

Add 500 µl (10 mM Tris.Cl pH8.0, 10 mM EDTA, 100 mM NaCl, 2% (w/v) SDS, 40 mM DTT, filter sterilized) to the sample. Add 1.25 µl 20 mg/ml proteinase K (Boehringer) Incubate at 55°C for 2 hours, vortexing once or twice. Perform 2x 0.5 mL 1:1 phenol:$CHCl_3$ extractions. After each extraction, centrifuge 12,000 rpm 5 min in a microfuge and recover 0.4 mL supernatant. Add 35 µl NaAc pH5.2 plus 1 mL EtOH. Place sample on ice 45 min; then centrifuge 12,000 rpm 30 min, rinse, air dry 30 min, and resuspend in 100 µl TE.

3. PCR

PCR is performed in a mixture containing, per reaction: 1 µl genomic DNA; 4 µl each primer (10 pmol/µl stocks); 4 µl 10 x PCR buffer (100 mM Tris.Cl pH8.5, 500 mM KCl, 15 mM $MgCl_2$); 4 µl 2 mM dNTPs (made from 100 mM dNTP stocks); 1 U Taq polymerase (Perkin Elmer, 5 U/µl); H_2O to 40 µl. About 40 cycles (94°C 30 sec, 55°C 30 sec, 72°C 30 sec) are performed, but cycling conditions may need to be varied. These conditions are for 0.2 mL thin wall tubes in Perkin Elmer 9600. For products in the 200 to 1000 bp size range, check 2 µl of the reaction on a 1.5% 0.5xTBE agarose gel using an appropriate size standard. For larger or smaller volumes (20 - 100 µl), one can use the same amount of genomic DNA but adjust the other ingredients accordingly.

4. *In vitro* transcription

Mix: 3 µl PCR product; 4 µl 5x buffer; 2 µl DTT; 2.4 µl 10 mM rNTPs (100 mM solutions from Pharmacia); 0.48 µl 10 mM fluorescein-UTP (Fluorescein-12-UTP, 10 mM solution, from Boehringer Mannheim); 0.5 µl RNA polymerase (Promega T3 or T7 RNA polymerase); and add H_2O to 20 µl. Incubate at 37°C for 3

h. Check 2 μl of the reaction on a 1.5% 0.5xTBE agarose gel using a size standard. 5x buffer is 200 mM Tris pH 7.5, 30 mM $MgCl_2$, 10 mM spermidine, 50 mM NaCl, and 100 mM DTT (supplied with enzyme). The PCR product needs no purification and can be added directly to the transcription mixture. A 20 μl reaction is suggested for an initial test experiment and hybridization; a 100 μl reaction is considered "preparative" scale (the reaction can be scaled up to obtain more target). The amount of PCR product to add is variable; typically a PCR reaction will yield several picomoles of DNA. If the PCR reaction does not produce that much target, then one should increase the amount of DNA added to the transcription reaction (as well as optimize the PCR). The ratio of fluorescein-UTP to UTP suggested above is 1:5, but ratios from 1:3 to 1:10 - all work well. One can also label with biotin-UTP and detect with streptavidin-FITC to obtain similar results as with fluorescein-UTP detection.

For nondenaturing agarose gel electrophoresis of RNA, note that the RNA band will normally migrate somewhat faster than the DNA template band, although sometimes the two bands will comigrate. The temperature of the gel can effect the migration of the RNA band. The RNA produced from *in vitro* transcription is quite stable and can be stored for months (at least) at -20°C without any evidence of degradation. It can be stored in unsterilized 6xSSPE 0.1% triton X-100 at -20°C for days (at least) and reused twice (at least) for hybridization, without taking any special precautions in preparation or during use. RNase contamination should of course be avoided. When extracting RNA from cells, it is preferable to work very rapidly and to use strongly denaturing conditions. Avoid using glassware previously contaminated with RNases. Use of new disposable plasticware (not necessarily sterilized) is preferred, as new plastic tubes, tips, etc., are essentially RNase free. Treatment with DEPC or autoclaving is typically not necessary.

5. Fragmentation

Heat transcription mixture at 94 degrees for forty min. The extent of fragmentation is controlled by varying Mg^{2+} concentration (30 mM is typical), temperature, and duration of heating.

6. Hybridization, Scanning, & Stripping

A blank scan of the slide in hybridization buffer only is helpful to check that the slide is ready for use. The buffer is removed from the flow cell and replaced with 1 mL of (hydrolysed) RNA in hybridization buffer and mixed well. Incubate for 15 - 30 min at 18°C. Remove the hybridization solution, which can be saved for subsequent experiments. Rinse the flow cell 4 - 5 times with fresh changes of 6 x SSPE / 0.1% Triton X-100, equilibrated to 18°C. The rinses can be performed rapidly, but it is important to empty the flow cell before each new rinse and to mix the liquid in the cell thoroughly. A series of scans at 30 min intervals using a hybridization temperature of 25°C yields a very clear signal, usually in at least 30 min to two hours, but it may be desirable to hybridize longer, *i.e.*, overnight. Using a laser power of 50 μW and 50 μm pixels, one should obtain maximum counts in the range of hundreds to low thousands/pixel for a new slide. When finished, the slide can be stripped using warm water.

These conditions are illustrative and assume a probe length of ~15 nucleotides. The stripping conditions suggested are fairly severe, but some signal may remain on the slide if the washing is not stringent. Nevertheless, the counts remaining after the wash should be very low in comparison to the signal in presence of target RNA. In some cases, much gentler stripping conditions are effective. The lower the hybridization temperature and the longer the duration of hybridization, the more difficult it is to strip the slide. Longer targets may be more difficult to strip than shorter targets.

7. Amplification of Signal

A variety of methods can be used to enhance detection of labelled targets bound to a probe on the array. In one

embodiment, the protein MutS (from E. coli) or equivalent proteins such as yeast MSH1, MSH2, and MSH3; mouse Rep-3, and Streptococcus Hex-A, is used in conjunction with target hybridization to detect probe-target complex that contain mismatched base pairs. The protein, labeled directly or indirectly, can be added to the chip during or after hybridization of target nucleic acid, and differentially binds to homo- and heteroduplex nucleic acid. A wide variety of dyes and other labels can be used for similar purposes. For instance, the dye YOYO-1 is known to bind preferentially to nucleic acids containing sequences comprising runs of 3 or more G residues.

8. Detection of Repeat Sequences

In some circumstances, *i.e.*, target nucleic acids with repeated sequences or with high G/C content, very long probes are sometimes required for optimal detection. In one embodiment for detecting specific sequences in a target nucleic acid with a DNA chip, repeat sequences are detected as follows. The chip comprises probes of length sufficient to extend into the repeat region varying distances from each end. The sample, prior to hybridization, is treated with a labelled oligonucleotide that is complementary to a repeat region but shorter than the full length of the repeat. The target nucleic is labelled with a second, distinct label. After hybridization, the chip is scanned for probes that have bound both the labelled target and the labelled oligonucleotide probe; the presence of such bound probes shows that at least two repeat sequences are present.

While the foregoing invention has been described in some detail for purposes of clarity and understanding, it will be clear to one skilled in the art from a reading of this disclosure that various changes in form and detail can be made without departing from the true scope of the invention. All publications and patent documents cited in this application are incorporated by reference in their entirety for all

purposes to the same extent as if each individual publication or patent document were so individually denoted.

Mutation	Exon	Ex Size	Pop Freq	Location	Sequence Around Mutation Site	PRIMERS	Amp Sz
297-3 C>T	i2	109	Manchester	Sub C>T +3 Exon 3	CTTTTTATTCTTTTG(C>T)AGAGAATGGGATAGA	787/788	297
R75O	3	109	Manchester	Substitute G>A at 60	TAATGCCCTTCGGC(G>A)ATGTTTTTTCTGGA	787/788	297
300 del A	3	109	Manchester	Delete A at 4	ATTCTTTTGCAGAGAaTGGGATAGAGAGCTGGCT	787/788	297
E60X	3	109	Manchester	Substitute G>T at 14	GAATGGGATAGA(G>T)AGCTGGCTTCAAAGA	787/788	297
L88S	3	109	Manchester	Substitute T>C at99	CTATGGAATCTTTT(T>C)ATATTTAGGGGTAAG	787/788	297
G85E	3	109	0.70%	Substitute G>A at90	TTATGTTCTATG(G>A)AATCTTTTTATATTTAG	787/788	297
R117H	4	216	0.80%	Substitute G>A at 77	AACAAGGAGGAAC(G>A)CTCTATCGCGATTTAT	851/769	381
R117C	4	216	rare	Substitute C>T at 76	AACAAGGAGGAA(C>T)GCTCTATCGCGATTTAT	851/769	381
Y122X	4	216	0.30%	Substitute T>A at 93	TATCGCGATTTA(T>A)CTAGGCATAGGCTTATG	851/769	381
I148T	4	216	Fr Can (10%)	Substitute T>C at 170	GGCCTTCATCACA(T>C)TGGAATGCAGATGAGA	851/769	381
621+1G>T	i4	216	1.30%	Sub G>T after last base	GATTTATAAGAAG(G>T)TAATACTTCCTTGCAC	851/769	381
711+1G>T	i5	90	0.90%	Sub G>T after last base	CAAATTTGATGAA(G>T)TATGTACCTATTGATT	867/868	289
L206W	6a	164	Fr Can (10%)	Substitute T>G at 38	TGGATCGCTCCTT(T>G)GCAAGTGGCACTCCTC	934/935	331
1138 ins G	7	247	Manchester	Insert G at 137	AATCATCCTCCGGAAAgATATTCACCACCATCT	789/790	404
1154 ins TC	7	247	Manchester	Insert TC at 153	TATTCACCACCATCTCcATTCTGCATTGTT	789/790	404
1161 del C	7	247	Manchester	Delete C at 160	CCACCATCTCATTCTGcATTGTTCTGCGCATGG	789/790	404
R334W	7	247	0.40%	Substitute C>T at 131	AAGGAATCATCCTC(C>T)GGAAAATATTCATTA	789/790	404
R347H	7	247	0.10%	Substitute G>A at 171	CTGCATTGTTCTGC(G>A)CATGGCGGTCACTCG	789/790	404
R347L	7	247	rare	Substitute G>T at 171	CTGCATTGTTCTGC(G>T)CATGGCGGTCACTCG	789/790	404
R347P	7	247	0.50%	Substitute G>C at 171	CTGCATTGTTCTGC(G>C)CATGGCGGTCACTCG	789/790	404
1078 delT	7	247	1.10%	Delete T at 77	CTTCTTCTCAGGGTTCTTTGTGGTGTTTTTATC	789/790	404
1248+1 G>A	i7	247	Manchester	Sub G>A 1 after Exon 7	AAACAAAATACAG(G>A)TAATGTACCATAATG	789/790	404
A455E	9	183	0.40%	Substitute C>A at 155	AGGACAGTTGTTGG(C>A)GGTTGCTGGATCCA	891/892	386
G480C	10	192	rare	Substitute G > T at 45	GGAGCCTTCAGAG(G>T)GTAAAATTAAGCACA	760/850	304
Q493X	10	192	0.30%	Substitute C>T at 85	TCATTCTGTTCT(C>T)AGTTTTCCTGGATTAT	760/850	304
DI507	10	192	0.50%	Delete 126, 127, 128	ATTAAAGAAAATATcatCTTTGGTGTTTCCTATG	760/850	304
F508C	10	192	rare	Substitute T>G at 131	TAAAGAAAATATCATCT(T>G)TGGTGTTTCCTA	760/850	304
DF508	10	192	67.20%	Delete 129, 130, 131	ATTAAAGAAAATATCATctTGGTGTTTCCTATG	760/850	304
V520F	10	192	0.20%	Substitute G>T at 166	TAGATACAGAAGC(G>T)TCATCAAAGCATGCC	760/850	304
1717-1G>A	i10	95	1.10%	Sub G>A at+1 Ex11	TATTTTTGGTAATA(G>A)GACATCTCCAAGTTT	762/763	233
G542X	11	95	3.40%	Substitute G>T at 40	ACAATATAGTTCTT(G>T)GAGAAGGTGGAAT	762/763	233
S549N	11	95	rare	Substitute G>A at 62	AGGTGGAATCACACTGA(G>A)TGGAGGTCAACG	762/763	233
S549I	11	95	rare	Substitute G>T at 62	AGGTGGAATCACACTGA(G>T)TGGAGGTCAACG	762/763	233
S549R(A>C)	11	95	rare	Substitute A>C at 61	AGGTGGAATCACACTG(A>C)GTGGAGGTCAACG	762/763	233
S549R(T>G)	11	95	0.30%	Substitute T>G at 63	AGGTGGAATCACACTGAG(T>G)GGAGGTCAACG	762/763	233
G551D	11	95	2.40%	Substitute G>A at 68	ATCACACTGAGTGGAG(G>A)TCAACGAGCAAGA	762/763	233
G551S	11	95	rare	Substitute G>A at 67	ATCACACTGAGTGGA(G>A)GTCAACGAGCAAGA	762/763	233
Q552X	11	95	rare	Substitute C>T at 70	ACACTGAGTGGAGGT(C>T)AACGAGCAAGAATT	762/763	233
R553Q	11	95	rare	Substitute G>A at 74	TGAGTGGAGGTCAAC(G>A)AGCAAGAATTTCT	762/763	233
R553X	11	95	1.30%	Substitute C>T at 73	TGAGTGGAGGTCAA(C>T)GAGCAAGAATTTCTTT	762/763	233
A559T	11	95	rare	Substitute G>A at 91	GCAAGAATTTCTTTA(G>A)CAAGGTGAATAAC	762/763	233
R560T	11	95	0.40%	Substitute G>C at 95	AATTTCTTTAGCAA(G>C)GTGAATAACTAA	762/763	233
R560K	11	95	rare	Substitute G>A at 95	GAATTTCTTTAGCAA(G>A)GTGAATAACTAA	762/763	233
1898+1G>A	i12	95	0.90%	Sub G>A after last Ex12	GAAATATTTGAAAG(G>A)TATGTTCTTTGAAT	931/932	299
D648V	13	724	Nat Am (63%)	Substitute A>T at 177	AACTCATGGGATGTG(A>T)TTCTTTCGACCAAT	955/884	360
2184 del A	13	724	0.70%	Delete A at 286	GACAGAAACAAAAAAaCAATCTTTTAAACAGAC	955/884	360
2184 ins A	13	724	rare	Insert A after 286	GACAGAAACAAAAAAAaCAATCTTTTAAACAGAC	955/884	360
2789+5G>A	i14b	38	1.10%	Sub G>A 5 one after last	CTCCTTGGAAAGTGA(G>A)TATTCCATGTCCTA	885/886	374
3272-26A>G	i17a	228	rare	Sub A>G 26 before 17b	TTTATGTTATTTGCA(A>G)TGTTTTCTATGGAAA	782/901	414
3272-93T>C	i17a	228	rare	Sub T>C 93 before 17b	ATTTGTGATATGATTA(T>C)TCTAATTTAGTCTTT	782/901	414
R1066C	17b	228	rare	Substitute C>T at 57	AGGACTATGGACACTT(C>T)GTGCCTTCGGAGGGC	782/901	414
L1077P	17b	228	rare	Substitute T>C at 91	TTACTTTGAAACTC(T>C)GTTCCACAAAGCTC	782/901	414
Y1092X	17b	228	0.50%	Substitute C>A at 137	CCAACTGGTTCTTGTA(C>A)CTGTCAACACTGCG	782/901	414
M1101K	17b	228	Hut (65%)	Substitute T>A at 163	TGCGCTGGTTCCAAA(T>A)GAGAATAGAAATGAT	782/901	414
R1162X	19	249	0.90%	Substitute C>T at 16	ATGCGATCTGTGAGC(C>T)GAGTCTTTAAGTTC	784/785	356
3659 del C	19	249	0.80%	Delete C at 59	AAGGTAAACCTAcCAAGTCAACCAAACCATACA	784/785	356
3849+4 A>G	i19	249	1.00%	Sub A>G 4 after last base	TCCTGGCCAGAGGGTG(A>G)GATTTGAACACT	784/785	356
3849+10kb	i19	10kb	1.40%	Sub C>T EcoR1 Fragment	ATAAAATGG(C>T)GAGTAAGACA	792/791	450
W1282R	20	156	rare	Substitute T>C at 127	AATAACTTTGCAACAG(T>C)GGAGGAAAGCCTTT	784/788	351
W1282X	20	156	2.10%	Substitute G>A at 129	AATAACTTTGCAACAGTG(G>A)AGGAAAGCCTTT	784/788	351
3905insT	20	156	2.10%	Insert T at 56	CTTTGTTATCAGCTTTTTTTGAGACTACTGAACAC	784/788	351
4005+1 G>A	i20	156	Manchester	Sub G>A after Exon 20	AGTGATACCACAG(G>A)TGAGCAAAAGGACTT	784/788	351
N1303K	21	90	1.80%	Substitute C>G at 36	CATTTAGAAAAAA(C>G)TTGGATCCCTATGAAC	766/793	396
N1303H	21	90	rare	Substitute A>C at 34	CATTTAGAAAA(A>C)ACTTGGATCCCTATGAAC		

Table 6